DOCUMENT RESUME

ED 222 561 TM 820 716

AUTHOR	Hambleton, Ronald K.; And Others
TITLE	Applications of Item Response Models to NAEP Mathematics Exercise Results.
INSTITUTION	Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.; Massachusetts Univ., Amherst. Laboratory of Psychometric and Evaluative Research.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	15 Feb 82
GRANT	NIE-G-80-0003
NOTE	238p.; Appendix B is marginally legible due to small print; For related documents, see TM 820 707-712.
EDRS PRICE	MF01/PC10 Plus Postage.
DESCRIPTORS	*Data Analysis; *Educational Assessment; Elementary Secondary Education; Equated Scores; Evaluation Methods; *Goodness of Fit; Item Analysis; Item Banks; *Latent Trait Theory; Mathematics Achievement; National Surveys; *Quantitative Tests; Test Construction; Test Items; Test Validity
IDENTIFIERS	National Assessment of Educational Progress; *NIE ECS NAEP Item Development Project; Second Mathematics Assessment (1978)

ABSTRACT

Item response model applications to National Assessment of Educational Progress (NAEP) data specifically aimed at the uses of item response models in mathematics item banking are discussed. Approaches for addressing goodness of fit were organized into three categories: Checks on model assumptions, expected features, and additional model predictions. Within the categories, several new methods were also advanced and several older methods which were not in common use for determining item response model-data fit were described. Many of these methods were then used to determine the fit of the one- and three-parameter models to six NAEP mathematics booklets (three booklets for 9-year-olds and three booklets for 13-year-olds) in the 1977-78 assessment. There were some inconsistent findings but it did appear that the three-parameter model provided an excellent fit to the data sets whereas the one-parameter model did not. When a bank of content valid and technically sound test items is available, and goodness of fit studies reveal a high match between the chosen item response model and the test data, item response models may be useful to NAEP in test development, detection of biased items, score reporting, equating test forms and levels, item banking and other applications as well. Primary type of information provided by the report: Results (Secondary Analysis). (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED222561

Applications of Item Response Models to NAEP Mathematics Exercise Results[1]

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

● Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Ronald K. Hambleton
-Principal Investigator-

Linda Murray
Robert Simon
-Research Assistants-

University of Massachusetts
Laboratory of Psychometric and Evaluative Research
Hills South, Room 152
Amherst, MA 01003

- February 15, 1982 -

TM 820 716

---

[1] The work upon which this publication is based was performed pursuant to ECS Contract No. 02-81-20319 which was issued under a jointly sponsored project of the Educational Commission of the States (ECS) and the National Institute of Education (NIE). It does not, however, necessily reflect the views of ECS or NIE.

Applications of Item Response Models to NAEP Mathematics Exercise Results

Ronald K. Hambleton, Principal Investigator
University of Massachusetts, Amherst

Abstract

In view of the technical advances and applications of item response models around the country, it is reasonable to expect ECS to consider the potential of item response models for use within its assessment programs. Among the areas to which the models could be applied are:

1. test building (item analysis, item bias, and item selection),
2. equating test forms,
3. measuring achievement growth,
4. linking NAEP exercises to other national, state, and district tests and test score norms.

But, the advantages derived from the applications of item response models cannot be achieved if the fit between the item response model of interest and NAEP exercises is less than adequate. Unfortunately, relatively little work has been done on the problem of determining the goodness of fit between an item response model and a test data set. Also, no one has looked at the fit between any of the item response models and NAEP mathematics exercises.

The research study had two principal objectives:

1. To organize and evaluate many of the available approaches for addressing the fit between an item response model and a data set.

2. To fit the one- and three-parameter logistic models to several NAEP mathematics exercise booklets, and evaluate and compare the results.

Approaches for addressing goodness of fit were organized into three categories: Checks on model assumptions, expected features, and additional model predictions. Within the categories, several new methods were also advanced and several older methods which were not in common use for determining item response model-data fit were described. Many of these methods were then used to determine the fit of the one- and three-parameter models to six NAEP mathematics booklets (three booklets for nine year olds and three booklets for thirteen year olds) in the 1977-78 assessment. There were some inconsistent findings but it did appear that the three-parameter model provided an excellent fit to the data sets whereas the one-parameter model did not. Recommendations for conducting future goodness of fit investigations were offered in the final section of the report.

Table of Contents

Abstract i

1.0 Introduction 1

- 1.1 Statement of Problems 1
- 1.2 Objectives 5
- 1.3 Item Response Models, Assumptions, Basic Concepts 6

2.0 Goodness of Fit Approaches 13

- 2.1 Overview 13
- 2.2 Statistical Tests of Significance 14
- 2.3 Checking Model Assumptions 18
- 2.4 Checking Model Features 27
- 2.5 Checking Additional Model Predictions 29
- 2.6 Summary 36

3.0 Analysis of NAEP Mathematics Exercises 38

- 3.1 Introduction 38
- 3.2 Description of NAEP Mathematics Exercises 38
- 3.3 Description of Data 41
- 3.4 Checking Model Assumptions 59
- 3.5 Checking Model Features 65
- 3.6 Checking Additional Model Predictions 93

4.0 Conclusions 175

- 4.1 Implications of Findings for NAEP 175

5.0 References 178

Appendix A - Item Response Model Goodness of Fit Studies

Appendix B - Item Response Model Residual Analysis Program
(Program Listing and Sample Output)

1.0 Introduction

1.1 Statement of Problems

Item response theory, or latent trait theory as it has sometimes been called, is the most popular topic for research at the present time among measurement specialists. There are numerous published research studies and conference presentations, and plentiful and diverse applications of the theory (for example, see Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1978; Lord, 1980). At least six books on the topic are in preparation; Applied Psychological Measurement will devote a special issue to the topic in 1982; and the Educational Research Institute of British Columbia will publish a special monograph in 1982 on promising item response model applications.

Presently, item response theory (IRT) is used by nearly all of the large test publishers, and many state departments of education and industrial and professional organizations to construct tests, to study item bias, to equate tests, and to report test score information. The many applications appear to be so successful that discussions of IRT have shifted from consideration of their advantages and disadvantages compared to classical test models to a consideration of topics such as model selection, item and ability parameter estimation, and methods for determining goodness of fit. Nevertheless, it would be incorrect to convey the impression that issues and technology associated with item response theory are fully developed and without controversy. Still, considerable progress has been made since the seminal papers by Lord (1952, 1953).

In view of the technical advances and applications of item response models (IRMs) it is reasonable to expect that the Educational Commission for the States (ECS) will consider in the near future the potential of IRMs for use within its assessment programs. Among the areas to which the models could be applied are:

1. in test building (item analysis, item bias, and item selection);
2. in equating test forms;
3. in reporting test scores;
4. in measuring achievement growth (on various groups of examinees, or on particular test items);
5. in assessing test score reliability;
6. in linking NAEP exercises to other national, state, and district test score norms.

It must be recognized however that any advantages derived from the applications of IRMs cannot be achieved if the fit between an IRM and a test data set of interest is less than adequate. Unfortunately, to date, relatively little research has addressed the problem of determining the goodness of fit between an IRM and a test data set of interest. What work has been done involves statistical tests, but these tests cannot be used as the sole determiner of model-data fit because of their dependence on examinee sample size. When sample sizes are large (as they will be with NAEP test data), nearly all departures between a model and a data set (even those where the practical significance of the difference is minimal) will lead to rejection of the null hypothesis of model-data fit. With small sample sizes even big differences may not be detected via statistical methods because of the low level of statistical power. ECS and others interested in applying IRMs could benefit from a set of recommendations for addressing goodness-of-fit studies. Unfortunately, the extant literature has not been compiled or organized, nor, to our knowledge, has much of the literature been critically evaluated.

Another problem is that the fit between any of the IRMs and NAEP mathematics exercises has not been studied. Of special interest in this study are goodness-of-fit results pertaining to several of the more promising applications of IRMs with the NAEP exercises. One of these applications involves creating an item bank with released items and then "linking" all of the items at a given age level to a common ability scale. Non-NAEP items can also be calibrated and added to the bank. In theory, statistical descriptors of items (item parameters) that are obtained from item response model analyses do *not* depend upon the choice of examinee groups used in estimating them, and expected ability estimates for examinees do *not* depend upon the particular choice of items selected from the bank. Such a system would permit, for example, schools to measure academic growth even though different test items are used at each test administration. Also, it would be possible to predict how well groups of examinees would have done on selected NAEP mathematics exercises (and comparisons can be made to the reported NAEP item norms) from their performance on other test items included in the item bank. Why would anyone wish to administer a different set of test items from those items which were normed? One reason is that teachers may wish to administer particular items to examinees because of their diagnostic value. A second reason is that with students who may be expected to do rather poorly or well on a test, better estimates of their abilities can be obtained when test items are selected to match their expected ability levels (Hambleton, 1979). There are other uses of item banks as well (see, for example, Hambleton et al., 1978). Again, however, these desirable outcomes will only be obtained if there is a more than adequate fit for an item response model to the NAEP mathematics exercise data. Of special interest is the invariance of item parameter estimates. For example, when items function differently for males and females; blacks, hispanics, and whites; and students from computationally-oriented and non-computationally-oriented math programs; IRM assumptions are violated and desired outcomes are not achieved. It is important to determine to what extent invariance of item parameters is obtained and over which sub-populations of examinees because the findings from item invariance studies have a direct bearing on the utility of IRMs in item banking. That is, when item statistics are not invariant, the usefulness of the item statistics, norms, etc., associated with an item bank are limited.

In summary, it would appear that there are several reasons for ECS to consider the utility of IRMs in their test development, analysis, and score reporting work. However, some preliminary work on approaches for assessing goodness of fit must be done first. With the approaches in hand, a variety of goodness-of-fit studies can be conducted on the NAEP mathematics exercises. Finally, at this time the advantages and disadvantages of the one- and three-parameter logistic models in relation to the NAEP exercises is unknown. Some work in the area would help ECS select the proper model, if they decide to use IRMs in one or more aspects of their testing methods and procedures.

1.2 Objectives

The research study had two principal objectives:

1. To organize and evaluate many of the available approaches for addressing the fit between an item response model and a data set.
2. To fit the one- and three-parameter logistic models to several NAEP mathematics exercise booklets, and evaluate and compare the results.

The potential of item response theory for solving a variety of NAEP testing and measurement problems appears to be substantial. However, this promise or potential is not guaranteed by simply processing test results through an available computer program to perform item response model analyses. Also, it cannot be assumed that because so many other data sets have been fit by item response models that the fit to NAEP exercise data is assured. The fact is that many of the applications described in the literature and especially the large set of AERA, NCME, and NAEP conference papers have failed to adequately address the goodness-of-fit issue and so the extent of model-data fit is unknown. Also, because of the national importance and visibility of NAEP, it is essential to carefully evaluate any proposed changes or additions to NAEP's approaches for building exercises and reporting and using the test information. Presently, NAEP is very successful, highly visible, and important. Therefore, there is no reason to take risks in test development and score reporting. In this research project, ECS is provided with a framework and methods for addressing the goodness-of-fit question. And, the work in this area should impact on other groups who are interested in addressing model-data fit questions. Second, ECS is provided with information pertaining to the fit between several NAEP mathematics exercise booklets and the one- and three-parameter logistic test models.

1.3 Item Response Models, Assumptions, Basic Concepts[1]

For many years now the classical test model has been useful to test developers and test score users. The model is based on "weak assumptions" and therefore the model can be applied to many testing problems (Lord & Novick, 1968). But, in spite of the wide acceptance of the classical test model, it has several important limitations. One limitation is that the two most common classical descriptors of test items, item difficulty and item discrimination, vary as a function of the average ability and the range of ability found in the particular sample of examinees for which they are computed. The usefulness of these item statistics in building tests is limited therefore to groups similar to those from which the examinee sample was drawn. Sample-dependent item statistics are a serious handicap for test developers.

Another shortcoming is that examinee test scores depend upon the particular selection of items included in a test. If distinct samples of test items are drawn from a pool of items all designed to measure the same knowledge and skills, and these item samples differ in difficulty, the test scores an examinee can expect to earn on these samples will also differ. With item-dependent ability estimates, comparisons among examinees are limited to situations where examinees have been administered identical (or "parallel") sets of test items.

---

[1] The material in this section of the report was edited from a paper by Hambleton (1979).

A third shortcoming of the classical test model is that it assumes that the errors of measurement are the same for all examinees. It is not uncommon to observe, however, that some examinees perform tasks more consistently than others and that consistency varies with ability. Needed are test models which can provide information about the precision of test scores and that are free to vary from one test score to another.

Because of the shortcomings of the classical test model, psychometricians have been investigating and developing more appropriate test models. Considerable attention is being directed currently toward the field of latent trait theory, sometimes referred to as item response theory or item characteristic curve theory (Lord, 1980).

In a few words, item response theory postulates that (1) underlying examinee performance on a test is a single ability or trait, and (2) the relationship between examinee performance on each item and the ability measured by the test can be described by a monotonically increasing curve. The curve is called an item characteristic curve and it provides the probability of examinees at various ability levels answering an item correctly. In Figure 1.3.1 below, two item characteristic curves are shown.

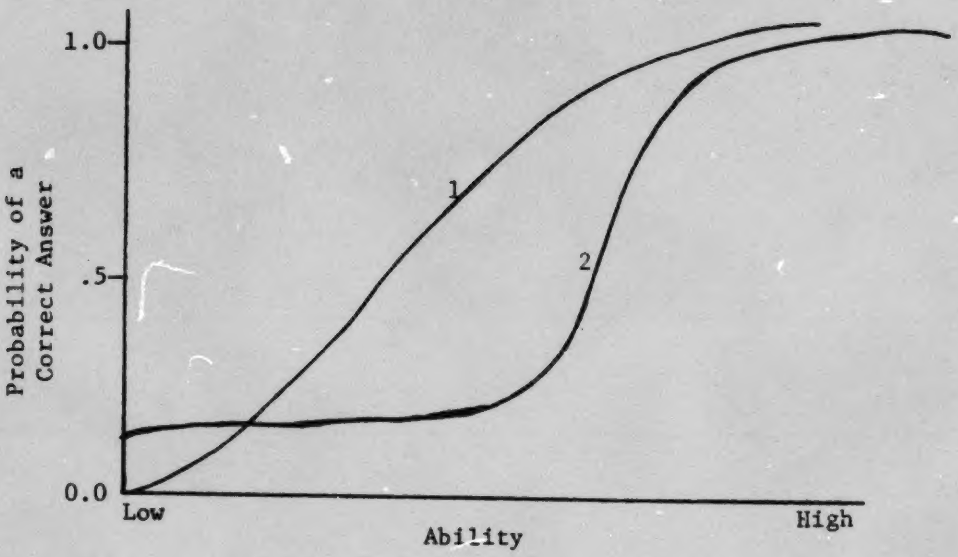


Figure 1.3.1. Two item characteristic curves

It is clear from the figure that the probability of a correct answer depends on the level of examinee ability. Examinees with more ability have higher probabilities for giving correct answers to items than lower ability examinees. Item characteristic curves are typically described by one-, two-, and three-parameter curves. The three item parameters are called item difficulty, item discrimination, and item pseudo-chance level. Items which are shifted to the right end of the ability scale are more difficult than those shifted to the left end of the ability scale. It is clear from Figure 1.3.1 then that item 2 is more difficult than item 1. The slope of an item characteristic curve describes an item's discriminating power. In Figure 1.3.1, therefore, item 2 is more discriminating than item 1. Finally, the probability of a very low ability examinee answering an item correctly is the item's pseudo-chance level. With item 1 in the figure, the probability is 0. With item 2 the probability is somewhat higher.

Most item response models, and all of the models which are presently popular, require the assumption that the test items are homogeneous in the sense that they measure a single ability or trait. In addition, it is common to assume that the item characteristic curves are described by one-, two-, or three-parameters, and the corresponding models are referred to as one-, two-, and three parameter models, respectively. With the three-parameter model, items can vary in their difficulty, discrimination level, and pseudo-chance level. With the two-parameter model, the pseudo-chance level parameter is 0 for all items. With the one-parameter model, not only does the pseudo-chance level parameter have a value of 0 for all items, but all items have a common level of discrimination.

When the assumptions of item response theory can be met in the data sets to which it is applied, at least a reasonable degree, what is obtained are (1) examinee ability estimates in the pool of items from which the items are drawn that do not depend upon the particular sample of items selected for the test, (2) item descriptors or statistics (difficulty, discrimination, pseudo-chance level) that do not depend upon the particular sample of examinees from the population of examinees for whom the earlier mentioned item pool is suitable, and (3) a statistic is provided indicating the precision with which each examinee's ability is estimated. Of course, the extent to which the three advantages are gained in an application of an item response model depends upon the closeness of the "fit" between a set of data and the model. If the fit is poor, the three desirable features either will not be obtained or obtained in a low degree (Lord, 1980).

Item response models are based on a set of assumptions about the test data. Two of these assumptions will be discussed here: Dimensionality, and the mathematical form chosen for the item characteristic curves. With respect to dimensionality, it is common to assume that only one ability is necessary to "explain," or "account" for examinee test performance. Item response models in which a single latent ability is presumed sufficient to explain or account for examinee performance are referred to as unidimensional models. The assumption that a set of test items is unidimensional is commonly made because scores on tests that measure only one trait are relatively easy to interpret. There exists no well accepted method for studying the unidimensionality of a set of test items. Factor analysis is the most common of the psychometric approaches used to address the dimensionality question (Hambleton et al., 1978).

An item characteristic curve (ICC) is a mathematical function that relates examinee probability of success on an item to the ability measured by the set of items contained in the test. $P_i(\theta)$ designates the probability of a correct response to item i by an examinee with ability level θ. The main difference to be found among currently popular latent trait models is in the mathematical form of the ICCs.

Birnbaum (1968) proposed ICCs which take the form of two-parameter logistic functions:

$$P_i(\theta) = \frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}} \quad , \quad (i=1, 2, \ldots, n). \qquad [1.3.1]$$

In equation [1.3.1], $P_i(\theta)$ is the probability that an examinee with ability θ answers item i correctly, and b_i and a_i are parameters of item i. The parameter, b_i, is referred to as item difficulty. It is the point on the ability scale such that examinees who possess that amount of ability have a 50% chance of answering an item correctly. The parameter, a_i, called item discrimination, is proportional to the slope of $P_i(\theta)$ at the point $\theta = b_i$. The constant D is a scaling factor set equal to 1.7.

A three-parameter model can be constructed from the two-parameter model by adding a third parameter, denoted c_i. The form of the three-parameter logistic curve is

$$P_i(\theta) = c_i + (1-c_i)\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}}, \quad (i=1, 2, \ldots, n). \quad [1.3.2]$$

The parameter c_i is the lower asymptote of the ICC and gives the probability of low ability examinees correctly answering the item.

The one-parameter model (sometimes called the "Rasch model") is a special case of the three-parameter logistic model in which guessing behavior is minimal (c_i=0), all items are assumed to have equal discriminating power, and items vary only in terms of difficulty. Therefore,

$$P_i(\theta) = \frac{e^{\theta-b_i}}{1+e^{\theta-b_i}} \quad . \quad [1.3.3]$$

The scale on which ability estimates are located is arbitrary. The scale is chosen so that ICCs of the form specified by the model under investigation fit as closely as possible to the available test data. An assumption is made that the correct metric is the one which maximizes predictions between unobservable characteristics (ability and item parameters) and observable data (examinee item responses). Since both item difficulty indices and ability estimates are measured on a common scale, it is usual to set the mean and standard deviation of one of the two variables to 0 and 1, respectively. In fact, any linear transformation of ability scale units is permissible and predictions from the model will not be influenced so long as the item discrimination parameters are revised accordingly. This means, for example, that if an agency wanted scores from an instrument on a scale with mean ability = 100 and standard deviation = 10, then, ability scores and item difficulties must be transformed using the linear equations

$$\theta_a^* = 10\,\theta_a + 100$$

$$b_i^* = 10\,b_i + 100 \quad ,$$

and the values of a_i must be transformed by the equation

$$a_i^* = \frac{1}{10}\,a_i \quad .$$

This last equation is determined so that

$$a_i^*\,(\theta_a^* - b_i^*) = a_i\,(\theta_a - b_i) \quad .$$

If this were not the case, the predictions for the model would be influenced by a change in the ability scale (Hambleton, 1980).

2.0 Goodness of Fit Approaches[1]

2.1 Overview

Item response models offer a number of advantages for test score interpretations and reporting of NAEP results but the advantages will only be obtained in practice when there is a close match between the model selected for use and the test data.

From a review of the relevant literature it appears that the determination of how well a model accounts for a set of test data can be addressed in at least three ways:

a. Determine if the test data satisfy the assumptions of the test model of interest.

b. Determine if the expected advantages derived from the use of the item response model (for example, invariant item and ability estimates) are obtained.

c. Determine the closeness of the fit between predictions and observable outcomes (for example, test score distributions) utilizing model parameter estimates and the test data.

Strictly speaking, tests of model assumptions are not tests of goodness of fit but because of their centrol role in model selection and use in the interpretation of goodness of fit tests we have included them first in a series of desirable goodness of fit investigations.

Promising practical approaches for addressing each category above will be addressed in subsequent sections. First, however, the inappropriateness of placing substantial emphasis on results from statistical tests will be explained.

---

[1] Small sections of this chapter are from Hambleton et al (1978) and Hambleton (1980).

2.2 Statistical Tests of Significance

Statistical tests of goodness of fit of various item response models have been given by many authors (Andersen, 1973; Bock, 1972; Mead, 1976; Wright, Mead, & Draba, 1976; Wright & Panchapakesan, 1969; Wright & Stone, 1979). The procedure advocated by Wright and Panchapakesan (1969) for testing the fit of the one-parameter model is one of the most commonly used. It essentially involves examining the quantity f_{ij} where f_{ij} represents the frequency of examinees at the ith ability level answering the jth item correctly. Then, the quantity y_{ij}, where

$$y_{ij} = \{f_{ij} - E(f_{ij})\}/\{\text{Var } f_{ij}\}^{\frac{1}{2}}$$

is distributed normally with zero mean and unit variance. Since f_{ij} has a binomial distribution with parameter p_{ij}, the probability of a correct response is given by $\theta_i^*/(\theta_i^* + b_j^*)$ for the one-parameter model, and r_i, the number of examinees in the score group. Hence, $E(f_{ij}) = r_i p_{ij}$, and $\text{Var }(f_{ij}) = r_i p_{ij}(1-p_{ij})$. Thus a measure of the goodness of fit, χ^2, of the model can be defined as

$$\chi^2 = \sum_{i=1}^{n-1} \sum_{j=1}^{n} y_{ij}^2$$

The quantity, χ^2, defined above has been assumed by Wright and his colleagues to have a χ^2 distribution with degrees of freedom (n-1) (n-2) since the total number of observations in the matrix $F=\{f_{ij}\}$ is n(n-1), and the number of parameters estimated is 2(n-1). Wright and Panchapakesan (1969) also defined a goodness-of-fit measure for individual items as

$$\chi_j^2 = \sum_{i=1}^{n-1} y_{ij}^2$$

where χ_j^2 is assumed to be distributed as χ^2 with degrees of freedom, (n-2). This method for determining the goodness of fit can also be extended to the two- and three-parameter item response models although it has not been extended to date.

There are several problems associated with the chi-square tests of fit discussed above. The χ^2 test has dubious validity when any one of the $E(f_{ij})$ terms, $\underline{i} = 1, 2, \ldots, \underline{n} - 1$; $\underline{j} = 1, 2, \ldots, \underline{n}$, have values less than one. This follows from the fact that when any of the $E(f_{ij})$ terms are less than one, the deviates y_{ij}, $1 - 1, 2, \ldots, n - 1$; $j = 1, 2, \ldots, n$, are not normally distributed and a χ^2 distribution is obtained only by summing the squares of normal deviates. Another problem encountered in using the χ^2 test is that it is sensitive to sample size. If enough observations are taken, the null hypothesis that the model fits the data will always be rejected using the χ^2 test. Divgi (1981) and Wollenberg (1980, 1982a, 1982b) have also demonstrated that the Wright-Panchapakesan goodness-of-fit statistic is not distributed as a χ^2 variable and the associated degrees of freedom have been assumed to be higher than they actually are. Clearly there are substantial reasons for not relying on the Wright-Panchapakesan statistic because of the role sample size plays in its interpretation and because of questions concerning the appropriate sampling distribution and degrees of freedom.

Alternately, Wright, Mead, and Draba (1976) and Mead (1976) have suggested a method of test of fit for the one-parameter model which involves conducting an analysis of variance on the variation remaining in the data after removing the effect of the fitted model. This procedure allows not only a determination of the general fit of the data to the model but also enables the investigator to pin-point guessing as the major factor contributing to the misfit. This procedure for testing goodness of fit of the one parameter model involves computing residuals in the data after removing the effect of the fitted model. These residuals are plotted against $(\theta_i - b_g)$. According to the model, the plot should be represented by a horizontal line through the origin. For guessing, the residuals follow the horizontal line until the guessing becomes important. When this happens the residuals are positive since persons are doing better than expected and in that region have a negative trend. If practice or speed is involved, the items which are affected display negative residuals with a negative trend line over the entire range of ability. Bias for a particular group may be detected by plotting the residuals separately for the two groups. It is generally found that the residuals have a negative trend for the unfavored group and a positive trend for the favored group.

When maximum likelihood estimates of the parameters are obtained, likelihood ratio tests can be obtained for hypotheses of interest (Waller, 1981). Likelihood ratio tests involve evaluating the ratio, λ, of the maximum values of the likelihood function under the hypothesis of interest to the maximum value of the likelihood function under the alternate hypothesis. If the number of observations is large, $-2 \log \lambda$ is known to have a chi-square distribution with degrees of freedom given by the difference in the number of parameters estimated under the alternate and null hypotheses. An advantage possessed by likelihood ratio tests over the other tests discussed earlier is apparent. Employing the likelihood ratio criterion, it is possible to assess the fit of a particular latent trait model against an alternative.

Andersen (1973) and Bock and Liebermann (1970) have obtained likelihood ratio tests for assessing the fit of the Rasch model and the two-parameter normal ogive model respectively. Andersen (1973) obtains a conditional likelihood ratio test for the Rasch model based on the within score group estimates and the overall estimates of item difficulties. He shows further that -2 times the logarithm of this ratio is distributed as χ^2 with degrees of freedom, (n-1) (n-2). Based on the work of Bock and Liebermann (1970), likelihood ratio tests can be obtained for testing the fit of the two-parameter normal ogive model. It should be pointed out that these authors have obtained both conditional and unconditional estimates of the parameters. For the likelihood ratio test, it would be more appropriate if the unconditional model is used since with this model ability parameters are not estimated, and hence the likelihood ratio criterian can be expected to have the chi-square distribution. This procedure can be extended to compare the fits of one model against another (Andersen, 1973).

The major problem with this approach is that the test criteria are distributed as chi-square only asymptotically. But, as was mentioned earlier, when large samples are used to accommodate this fact, the chi-square value may become significant owing to the large sample size!

2.3 Checking Model Assumptions

Item response models are based on strong assumptions which will not be completely met by any set of test data (Lord & Novick, 1968). There is evidence that the models are robust to some departures but the extent of robustness of the models has not been firmly established (Hambleton et al., 1978). Given doubts of the robustness of the models, one might be tempted to simply fit the most general model since it will be based on the least restrictive assumptions. Unfortunately, the more general models are multi-dimensional (i.e., assume that more than one latent variable is required to account for examinee test performance), and they are complex and do not appear ready for wide-scale use. Alternately, it has been suggested that the three-parameter logistic model, the most general of the unidimensional models in common use, be adopted. In theory, the three-parameter model should result in better fits than either the one- or two-parameter models. But, there are three problems with this course of action: (1) more computer-time is required to conduct the analyses, (2) somewhat larger samples of examinees and items are required to obtain satisfactory item and ability estimates, and (3) the additional item parameters (item discrimination and pseudo-chance levels) complicate the use of the model for practitioners. Of course, in spite of the problems, and with important testing programs such as NAEP and a highly trained staff, the three-parameter model may still be preferred.

Model selection can be aided by an investigation of four principal assumptions of several of the item response models: unidimensionality, equal discrimination indices, minimal guessing, and non-speeded test administrations. Promising approaching for studying these assumptions are summarized in Figure 2.3.1 and will be briefly considered next.

Figure 2.3.1 Approaches for Conducting Goodness of Fit Investigations

Checking Model Assumptions

1. Unidimensionality (Applies to Nearly All Item Response Models)

   - Kuder-Richardson Formula 20 (Common Approach But Not Acceptable - Statistic is Influenced By Test Score Variability and Test Length).
   - Plot of Eigenvalues (From Largest to Smallest) of the Inter-Item Correlation Matrix - Look for a Dominant First Factor, and a High Ratio of the First to the Second Eigenvalue (Reckase, 1979).
   - Comparison of Two Plots of Eigenvalues - the One Described Above and One of Eigenvalues for an Inter-Item Correlation Matrix of Random Data (Same Sample Size, and Number of Variables, Random Data Normally Distributed) (Horn, Psychometrika, 1965).
   - Plot of Content-Based Versus Total-Test Based Item Parameter Estimates (Bejar, JEM, 1980).
   - Analysis of Residuals After Fitting a One Factor Model to the Inter-Item Covariance Matrix (McDonald, BJMSP, 1980).

2. Equal Discrimination Indices (Applies to the One-Parameter Logistic Model)

   - Analysis of Variability of Item-Test Score Correlations (For Example, Point-Biserial and Biserial Correlations).
   - Identification of Percent of Item-Test Score Correlations Falling Outside Some Acceptable Range (For Example, the Average Item-Test Score Correlation ± .15).

3. Minimal Guessing (Applies to the One- and Two-Parameter Logistic Model)

   - Investigation of Item-Test Score Plots (Baker, JEM, 1964, 1965).
   - Consideration of the Performance of Low-Ability Examinees (Selected with the Use of Test Results, or Instructor Judgments) on the Most Difficult Test Items.
   - Consideration of Item Format and Test Time Limits (For Example, Consider the Number of Item Distractors, and Whether or Not the Test Was Speeded).

4. Non-speeded (Power) Test Administration (Applies to Nearly All Item Response Models).

   - Comparison of Variance of the Number of Items Unattempted to the Variance of the Number of Items Answered Wrongly (Gulliksen, 1950).
   - Investigation of the Relationship Between Scores on a Test With the Specified Time Limit and With an Unlimited Time Limit (Cronbach and Warrington, 1951).

Figure 2.3.1 (continued)

- Investigation of (A) Percent of Examinees Completing the Test, (B) Percent of Examinees Completing 75% of the Test, and (C) Number of Items Completed by 80% of the Examinees (ETS Method, See Donlon, 1978).

Checking Expected Model Features

1. Invariance of Item Parameter Estimates (Applies to All Models)

   - Comparison of Item Parameter Estimates Obtained in Two or More Sub-groups of the Population for Whom the Test is Intended (For Example, Males and Females; Blacks, Whites, and Hispanics; Instructional Groups; High and Low Performers on the Test or Other Criterion Measure, Geographic Regions). Normally Comparisons Are Made of the Item Difficulty Estimates and Presented in Graphical Form (Scattergrams). Random Splits of the Population Into Sub-groups The Same Size Provide a Basis for Obtaining Plots Which Can Serve as a Baseline for Interpreting the Plots of Principal Interest. Graphical Displays of Distributions of Standardized Differences in Item Parameter Estimates Can Be Studied. Distributions Ought to Have a Mean of Zero and a Standard Deviation of One (For Example, Wright, 1968; Lord, 1980; Hambleton and Swaminathan, 1982).

2. Invariance of Ability Parameter Estimates (Applies to All Models)

   - Comparison of Ability Estimates Obtained in Two or More Item Samples From the Item Pool of Interest. Choose Item Samples Which Have Special Significance Such As Relatively Hard Versus Relatively Easy Samples, and Subsets Reflecting Different Content Categories Within the Total Item Pool. Again, Graphical Displays and Investigation of the Distribution of Ability Differences Are Revealing.

Checking Model Predictions of Actual (and Simulated) Test Results

- Investigation of Residuals and Standardized Residuals of Model-Test Data Fits at the Item and Person Levels. Various Statistics are Available to Summarize the Fit Information. Graphical Displays of Data Can Be Revealing.
- Comparison of Item Characteristic Curves Estimated in Substantially Different Ways (For Example, Lord, Psychometrika, 1970).
- Plot of Test Scores and Ability Estimates (Lord, Psychometrika, 1974).
- Plots of True and Estimated Item and Ability Parameters (For Example, Swaminathan, 1981; Hambleton and Cook, 1982). These Studies Are Carried Out With Computer Simulation Methods.
- Comparison of Observed and Predicted Score Distributions. Various Statistics (Chi-Square, For Example) and Graphical Methods Can Be Used to Report Results. Cross-Validation Procedures Should Be Used, Especially If Sample Sizes Are Small (Hambleton and Traub, BJMSP, 1973).
- Investigation of Hypotheses Concerning Practice Effects, Test Speededness, Cheating, Boredom, Item Format Effects, Item Order, etc.

Unidimensionality

The assumption of a unidimensional latent space is a common one for test constructors, since they usually desire to construct unidimensional tests so as to enhance the interpretability of a set of test scores (Lumsden, 1976). What does it mean to say that a test is unidimensional? Suppose a test consisting of n items is intended for use in r subpopulations of examinees (e.g., several ethnic groups). Consider next the conditional distributions of test scores at a particular ability level for the r subpopulations. These conditional distributions for the r subpopulations will be identical if the test is unidimensional. If the conditional distributions vary across the r subpopulations, it can only be because the test is measuring something other than the single ability. Hence, the test cannot be unidimensional.

It is possible for a test to be unidimensional within one population of examinees and not unidimensional in another. Consider a test with a heavy cultural loading. This test could appear to be unidimensional for all populations with the same cultural background. However, when administered to populations with varied cultural backgrounds, it may in fact have more than a single dimension underlying the test score. Examples of this situation are seen when the factor structure of a particular set of test items varies from one cultural group to another.

Lumsden (1961) provided an excellent review of methods for constructing unidimensional tests. He concluded that the method of factor analysis held the most promise. Fifteen years later he reaffirmed his conviction (Lumsden, 1976). Essentially, Lumsden recommends that a test constructor generate an initial pool of test items selected on the basis of empirical evidence and a priori grounds. Such an item selection procedure will increase the likelihood that a unidimensional set of test items within the pool of items can be found. If test items are not preselected, the pool may be too heterogeneous for the unidimensional set of items in the item pool to emerge. In Lumsden's method, a factor analysis is performed and items not measuring the dominant factor obtained in the factor solution are removed. The remaining items are factor analyzed, and again, "deviant" items are removed. The process is repeated until a satisfactory solution is obtained. Convergence is most likely when the initial item pool is carefully selected to include only items that appear to be measuring a common trait. Lumsden proposed that the ratio of first factor variance to second factor variance be used as an "index of unidimensionality."

Factor analysis can also be used to check the reasonableness of the assumption of unidimensionality with a set of test items (Hambleton & Traub, 1973). However, the approach is not without problems. For example, much has been written about the merits of using tetrachoric correlations or phi correlations (McDonald & Ahlawat, 1974). The common belief is that using phi correlations will lead to a factor solution with too many factors, some of them "difficulty factors" found because of the range of item difficulties among the items in the pool. McDonald and Ahlawat (1974) concluded that "difficulty factors" are unlikely if the range of item difficulties is not extreme and the items are not too highly discriminating.

Tetrachoric correlations have one attractive feature. A sufficient condition for the unidimensionality of a set of items is that the matrix of tetrachoric item intercorrelations has only one common factor (Lord & Novick, 1968). On the negative side, the condition is not necessary. Tetrachoric correlations are awkward to calculate (the formula is complex and requires some numerical integration), and, in addition, do not necessarily yield a correlation matrix that is positive definite, a problem when factor analysis is attempted.

Kuder-Richardson Formula 20 has on occasion been recommended and/or used to address the dimensionality of a set of test items. But Green, Lissitz, and Mulaik (1977) have noted that the value of KR-20 depends on test length and group heterogeneity and therefore the statistic provides misleading information about unidimensionality.

A somewhat more promising method involves considering the plots of eigenvalues for test item intercorrelation matrices and looking for the "breaks" in the plots to determine the number of "significant" underlying factors. To assist in locating a "break" Horn (1965) suggested that the plot of interest be compared to a plot of eigenvalues obtaining from an item intercorrelation matrix of the same size and where inter-item correlations are obtained by generating random variables from normal distributions. The same number of examinees as used in the correlation matrix of interest is simulated.

Another promising approach, in part because it is not based on the analysis of correlation coefficients, was suggested by Bejar (1980):

1. Split test items on an a priori basis (i.e., content considerations). For example, isolate a subset of test items which appear to be tapping a different ability from the remaining test items.

2. For items in the subset, obtain item parameter estimates twice: once by including the test items in item calibration for the total test and a second time by calibrating only the items in the subset.

3. Compare the two sets of item parameter estimates by preparing a plot (see Figure 2.3.2).

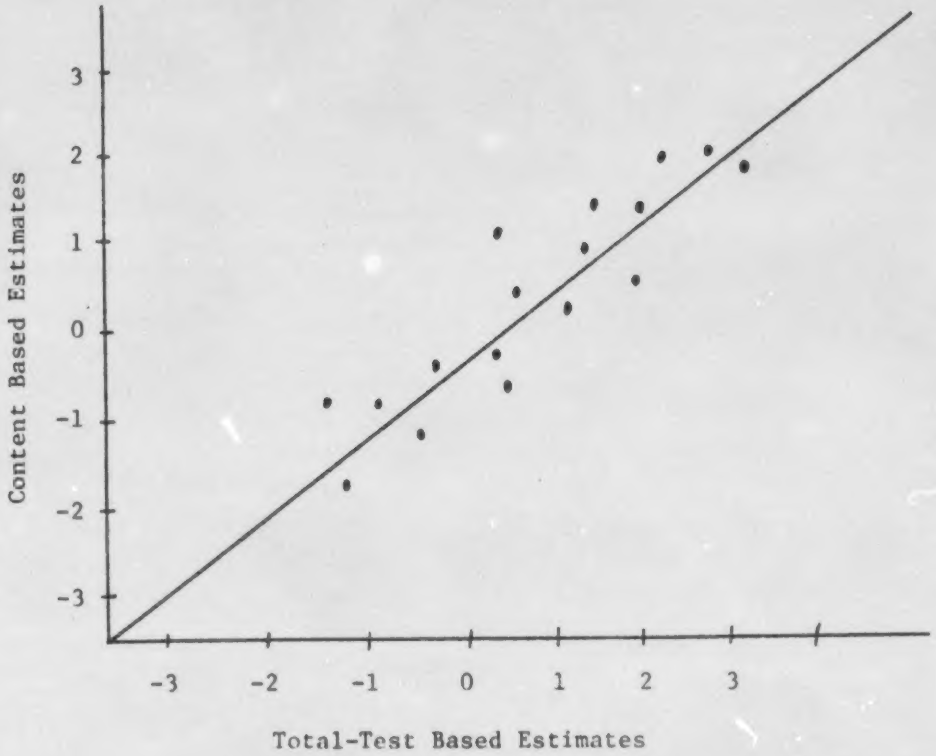
Unless the item parameter estimates (apart from sampling error) are equal, the probability for passing items for fixed ability levels will differ. This is not acceptable because it implies that performance on items depends on which items are included in the test which contradicts the unidimensionality assumption.

Finally, McDonald (1980a, 1980b) and Hattie (1981) have suggested the use of non-linear factor analysis and the analysis of residuals as a promising approach. The approach seems promising because test items are related to one another in a non-linear way anyway, and the analysis of residuals after fitting a one-factor solution seems substantially more revealing and insightful than conducting significance tests on the amount of variance accounted for.

Equal Discrimination Indices

This assumption is made with the one-parameter model. There appear to be only descriptive methods available for investigating departures from this model assumption. A rough check of its viability is accomplished by comparing the similarity of item point-biserial or biserial correlations. The range (or the standard deviation) of the discrimination indices should be small if the assumption is to be viable. Wright and his colleagues have, on occasion, looked at the residuals remaining after fitting a one-parameter model and attempted to study variation in item discrimination indices but they have written little on their methods.

Figure 2.3.2 Plot of content-based and total-test based item parameter estimates.



Guessing

There appears to be no direct way to determine if examinees guess the answers to items in a test. Two methods have been considered (1) non-linear item-test score regression lines, and (2) the performance of low test score examinees on the hardest test items. With respect to the first method, for each test item, the proportion of correct answers for each test score group (small test score groups can be combined to improve the accuracy of results) are plotted. Guessing is assumed to be operating when test performance for the low performing score groups exceeds zero. For method two, the performance of the low-scoring examinees on the hardest test questions if of central concern. Neither method however is without faults. The results will be misleading if the test items are relatively easy for the low ability group, and/or if the low ability group is only relatively low in ability in relation to other examinees in the population of examinees for whom the test is intended but now low ability in any absolute sense (i.e., very low scorers on the test).

Speededness of the Test

Little attention is given to this seldom stated assumption of many item response models. When it operates it introduces an additional factor influencing test performance. It can be identified by a factor analytic study. Interestingly, with some of the new ability estimation methods (Lord, 1980), the failure of examinees to complete a test can be handled so that the speededness factor does not "contaminate" ability score estimates. The appropriateness of the assumption in relation to a set of test results can be checked by determining the number of examinees who fail to finish a test and the number of items they fail to complete. The ideal situation occurs when examinees have sufficient time to attempt each question in a test.

Donlon (1978) provided an extensive review of methods for determining the speededness of tests. Three of the most promising are cited in

Figure 2.3.1. Perhaps discussion of only one here will suffice. It involves obtaining an estimate of the correlation between scores obtained under power and speed conditions and correcting the correlation for attenuation due to the unreliability associated with the power and speed scores:

$$\rho(T_p,T_s) = \frac{\rho(X_p,X_s)}{\sqrt{\rho(X_p,X_p')}\ \sqrt{\rho(X_s,X_s')}} \quad .$$

The speededness index proposed by Cronbach and Warrington (1951) is

$$\text{Speededness Index} = 1 - \rho^2(T_s,T_p) \quad .$$

The index is obtained in practice by administered parallel-forms of the test of interest under speed and power conditions to the same group of examinees.

2.4 Checking Model Features

When item response models fit test data sets, three advantages are obtained:

1. Examinee ability estimates are obtained on the same ability scale and can be compared even though examinees may have taken different sets of test items from the pool of test items measuring the ability of interest.
2. Item statistics are obtained which do not depend on the sample of examinees used in the calibration of test items.
3. An indication of the precision of ability estimates at each point on the ability scale is obtained.

It is to obtain the advantages that item response models are often chosen as the mode of analysis. However, whether or not these features are obtained in any application depends on many factors — model-data fit, test length, precision of the item parameter estimates, and so on. Through some fairly straightforward methods, these features can be studied and their presence in a given situation determined.

The first one can be addressed, for example, by administering examinees two or more samples of test items which vary widely in difficulty (Wright, 1968). In some instances, items can be administered in a single test and two scores for each examinee obtained: the scores are based on the easier and harder halves of the test. To determine if there is substantial difference in test difficulty, the distributions of scores on the two halves of the test can be compared. Pairs of ability estimates obtained from the two halves of the test for each examinee are plotted on a graph. The bivariate plot of ability estimates should be linear because expected ability scores for examinees do not depend upon the choice of test items when the item response model under investigation fits the test data. Some scatter of points about a best fitting line, however, is to be expected because of measurement error. When a linear relationship is not obtained, one or more of the underlying assumptions of the item response model under investigation are being violated by the test data set. Factors such as test characteristics, test lengths, precision of item statistics, and so on can also be studied to determine their influence.

The second feature is studied in essentially the same way as the first. The difference is that extreme ability groups are formed and item parameter estimates in the two samples are compared. Wright (1968) and Lord (1980) have carried out extensive studies in this area. Again, if the test data are fit by the item response model under investigation, there should be a linear relationship between item parameter estimates from the two examinee samples, even if the samples differ in ability, race, or sex (Lord & Novick, 1968). The comparison is carried out for each item parameter in the model of interest. With respect to NAEP exercises it seems especially important to compare item parameter estimates derived from (say) black and white examinee groups. This check would be a stiff one but a linear relationship must still be obtained or it must be said that the item response model does not fit the test data for one or two of the groups.

Perhaps the most serious weakness of the approaches described above (and these are the only ones found in the literature) is that there is no baseline data available for interpreting the plots. How is one to know whether the amount of scatter is appropriate, assuming model-data fit? Alternately, statistical tests are performed to study the differences between (say) b values obtained in two samples. But, as long as there is at least a small difference in the true parameter values in the samples, statistically significant differences will be obtained when sample sizes are large. Thus, statistically significant differences may be observed even when the practical differences are very small.

The third feature is a harder one to address. Perhaps it is best answered via simulation methods. According to the theory, if a test is "long enough," the conditional distribution of ability estimates at each ability level is normal (mean = ability; sd = $1/\sqrt{\text{information}}$). It appears that a test must include about 20 items (Samejima, 1977).

2.5. Checking Additional Model Predictions

Several approaches for checking model predictions were introduced in Figure 2.3.1. One of the most promising approaches for addressing model-data fit involves the use of residual analyses. An item response model is chosen; item and ability parameter estimates are obtained; and predictions of the performance of various ability groups on the items on the test are made, assuming the validity of the chosen model. Comparisons of the predicted results with the actual results are made.

By comparing the average item performance levels of various ability groups to the performance levels predicted by an estimated item characteristic curve, a measure of the fit between the estimated item characteristic curve and the observed data can be obtained. This process, of course, can and is repeated for each item in a test. In Figure 2.5.1, a plot of the residuals (difference between the observed data and an estimated item characteristic curve) across ability groups for four items are reported along with likely explanations for the results. The average item performance of each ability group is represented by the symbol "x" in the figure. If, for example, 25 of 75 examinees in the lowest ability group answered an item correctly, an "x" would be placed at a height of .33 above the average ability score in the ability group where the performance was obtained. (The width of each ability group should be wide enough to contain a reasonable number of examinees.) With items "a", "b", and "c" in Figure 2.5.1, there is substantial evidence of a misfit between the available test data and the estimated item characteristic curves (Hambleton, 1980). It is surprising to note, given their apparent usefulness, that residuals have not received more attention from item response model researchers.

Lord (1970, 1974) has advanced several approaches for addressing model-data fit. In 1970, Lord compared the shape of ICC curves estimated by different methods. In one method he specified the curves to be three-parameter logistic. In the other method no mathematical form of the ICCs was specified. Since the two methods gave very similar results (see Figure 2.5.2) he argued that it was reasonable to impose the mathematical form of three-parameter logistic curves on his data. Presumably Lord's study can be replicated on other data sets as well although his second

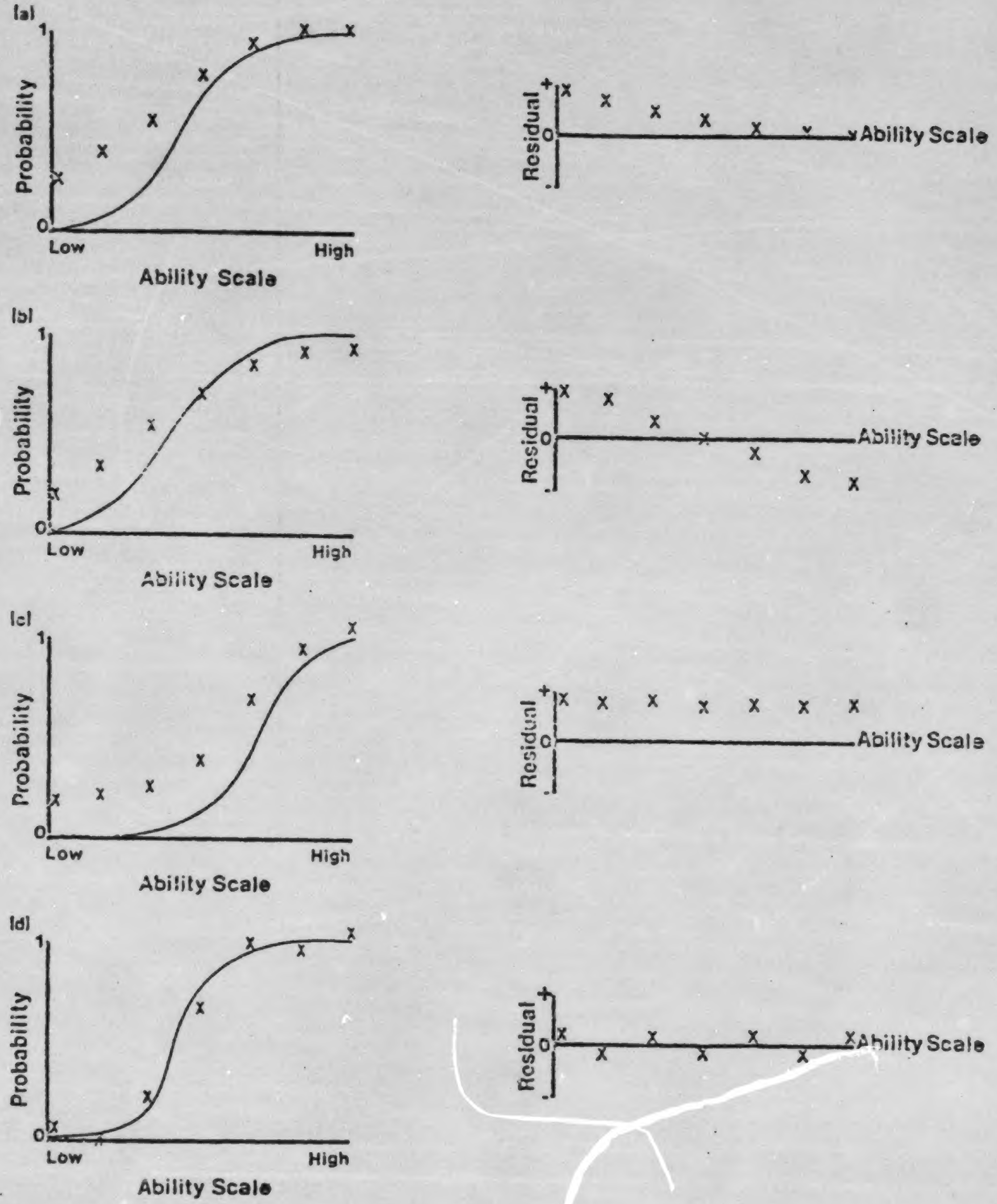


Figure 2.5.1. Analysis of residuals. Possible Explanation: (a) failure to account for "guessing", (b) failure to account for "item discrimination", (c) biased item, and (d) item fitted by the particular model.

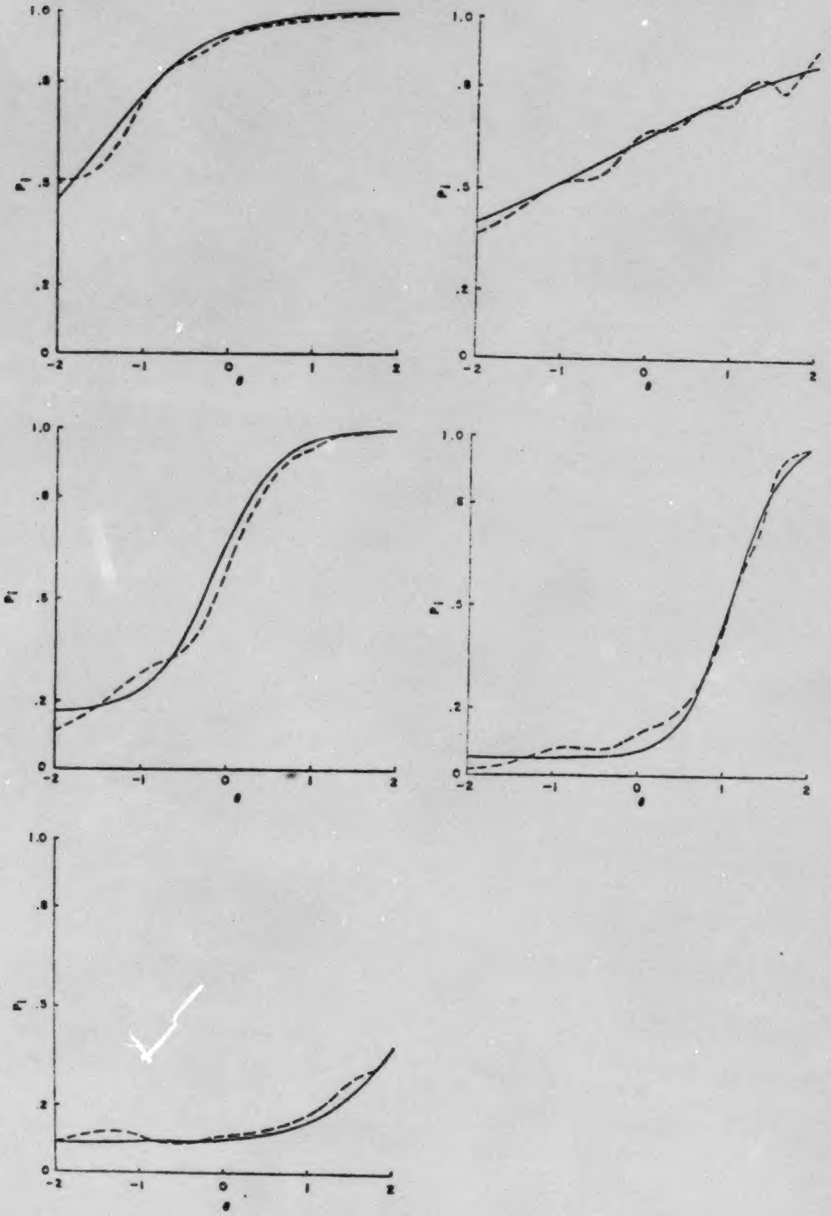


Figure 2.5.2. Five item characteristic curves estimated by two different methods (reproduced from Lord, 1970).

method required very large examinee samples. In a second study, Lord (1974) was able to assess, to some extent, the suitability of ability estimates by comparing them to raw scores. The relationship should be high but not perfect.

Simulation studies have been found to be of considerable value in learning more about item response models, and how they compare in different applications (e.g., Hambleton, 1982a, 1982b; Hambleton & Cook, 1982; Ree, 1979). It is possible to simulate data with known properties and see how well the models recover the true parameters. Hambleton and Cook (1982) found, for example, when concerned with estimating ability scores for ranking, description, or decisions, that the one-, two-, and three-parameter models provided highly comparable results except for low ability examinees. Swaminathan (1981) conducted a study of Bayesian estimators and used a comparison of true and estimated difficulty values to evaluate these procedures (see Figure 2.5.3).

Several researchers (for example, Hambleton & Traub, 1973; Ross, 1966) have studied the appropriateness of different mathematical forms of item characteristic curves by using them, in a comparative way, to predict test score distributions (See Figures 2.5.4 and 2.5.5). Hambleton and Traub (1973) obtained item parameter estimates for the one- and two-parameter models from three aptitude tests. Assuming a normal ability distribution and using test characteristic curves obtained from both the one- and two-parameter logistic models, they obtained predicted score distributions for each of the three aptitude tests. A χ^2 goodness of fit index was used to compare actual test score distributions with predicted test score distributions from each test model. Judgment can then be used to determine the suitability of any given test model and the desirability of one model

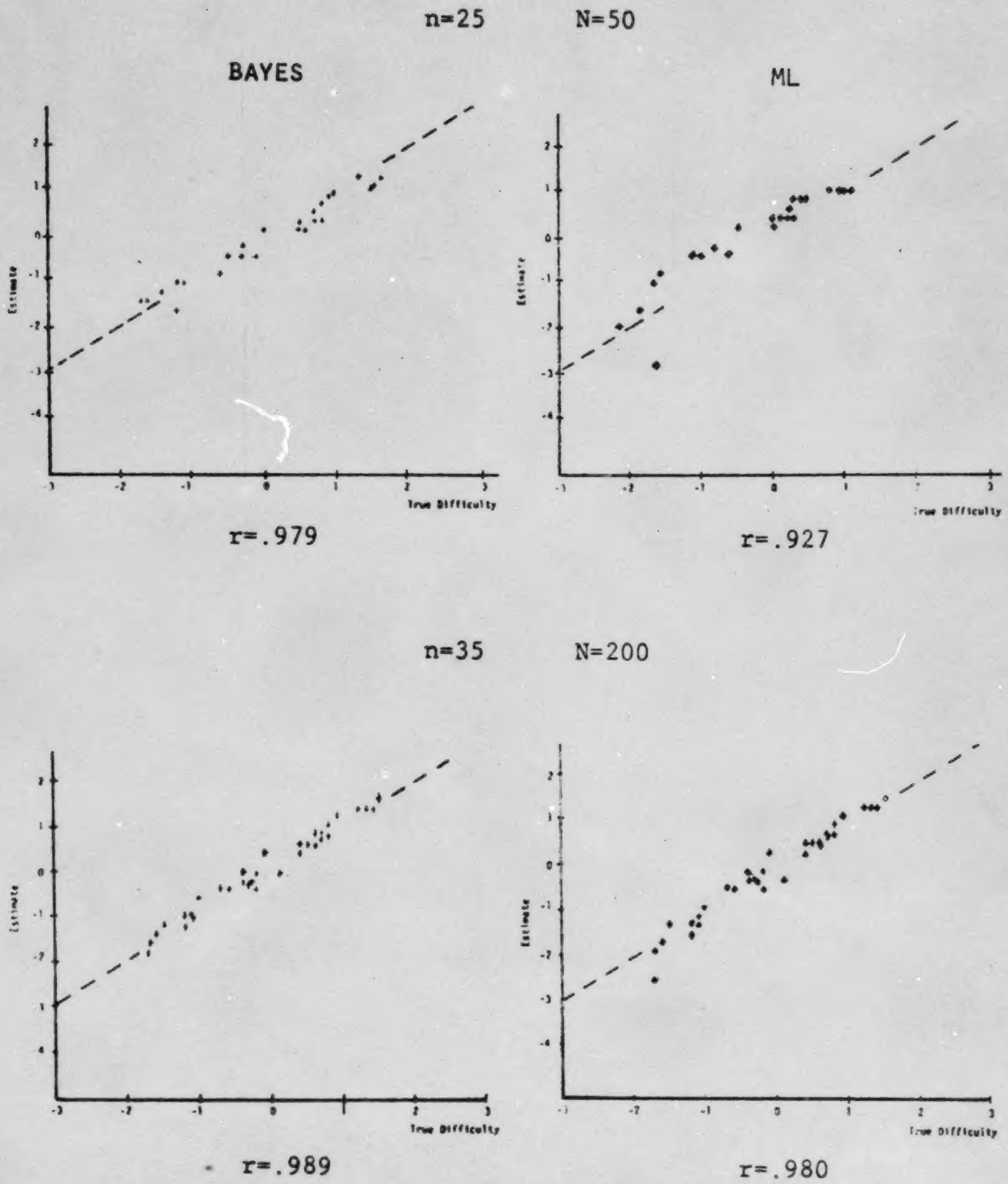


Figure 2.5.3. Bivariate plot of true and estimated values of difficulty (two-parameter model).

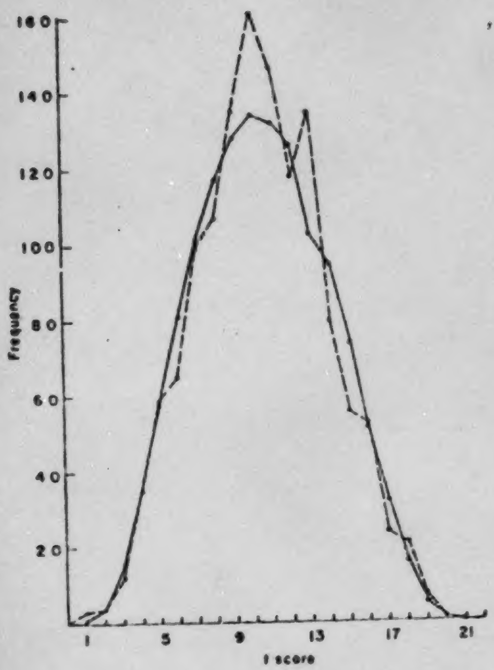


Figure 2.5.4. Observed (•) and expected (o) distributions for OSAT-Verbal using the two-parameter logistic model.

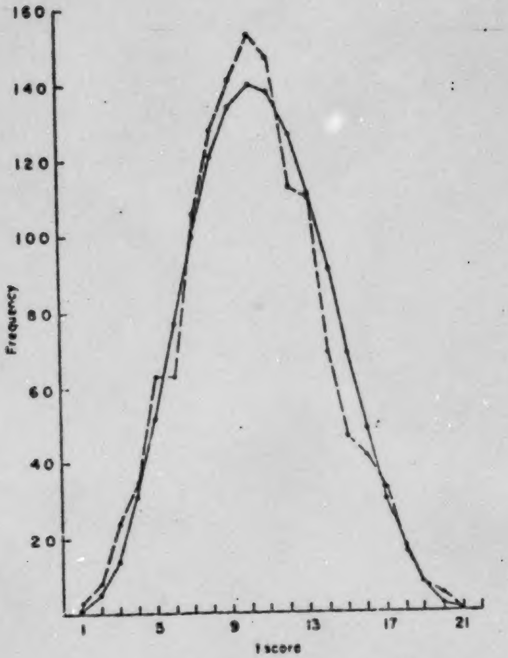


Figure 2.5.5. Observed (•) and expected (o) distributions for OSAT-Verbal using the one-parameter logistic model.

over another. Hambleton and Traub (1973) based their predictions upon a normal ability distribution assumption however it is neither desirable nor necessary to make such an assumption to obtain predicted score distributions.

Finally, it is reasonable and desirable to generate testable hypotheses concerning model-data fit. Hypotheses might be generated because they seem interesting (e.g., Are item calibrations the same for examinees receiving substantially different types of instruction?) or because questions may have arisen concerning the validity of the chosen item response model and testing procedure(e.g., What effect does the context in which an item is pilot-tested have on the associated item parameter estimates?) On this latter point, see for example, Yen (1980). Surprisingly, there is relatively little attention beyond the attention associated with category 1 and 2 for testing hypotheses.

2.6 Summary

Our review of relevant literature associated with conducting goodness of fit studies revealed a substantial number of approaches. From our perspective, however, there appeared to be too much emphasis on statistical tests for determining goodness of fit. As an alternative, the use of judgment in interpreting misfit statistics and other model-data comparisons for more than one model seems desirable. Perhaps the statistical approach can be replaced by the use of graphical methods, replications, cross validation techniques, study of residuals, baseline results to aid in interpretations, study of practical consequences of misfit, and so on.

With respect to testing model assumptions, unidimensionality is clearly the most important assumption to satisfy. Many tests of unidimensionality are available but those which are independent of correlations (Bejar) and/or incorporate the analysis of residuals (McDonald) seem most useful. In category two, there is a definite shortage of ideas and techniques. Presently, plots of (say) item parameter estimates obtained in two groups are compared but without the aid of any "baseline plots." Or, statistical tests are used to compare the two sets of item parameter estimates but such tests are less than ideal for reasons offered in section 2.2. Several new techniques seem possible and these will be introduced in the next chapter. In the third category, a number of very promising approaches have been described in the literature but they have received little or no attention from researchers. Perhaps the problem is due to a shortage of computer programs to carry out necessary analyses or to an over reliance on statistical tests. In any case the problem is likely to be overcome in the near future and we will focus our attention in the next chapter on several of the more promising approaches in this category.

3.0 Analysis of NAEP Mathematics Exercises

3.1 Introduction

In this section of the report (1) the NAEP mathematics exercises will be briefly described, (2) the particular mathematics exercises which were chosen for analysis will be described, and (3) the results from many item response model-NAEP math data fit investigations introduced in section 2 will be presented and discussed.

3.2 Description of NAEP Mathematics Exercises

In the 1977-78 NAEP assessment of mathematics skills of 9, 13, and 17 year olds, approximately 650 test items (called "exercises" by NAEP) at each age level were used. Available test items at a given age level were randomly assigned to one of ten forms. Each test form was administered to a carefully chosen sample of (approximately) 2500 examinees. Elaborate sampling plans were designed and carried out to insure that each form was administered to a nationally representative sample of examinees.

Item statistics play only a minor part in NAEP mathematics test development. Test items are included in test forms if they measure what national panels of mathematics specialists believe should be included in the NAEP testing program. Content considerations are dominant in the item selection process. In this respect test development parallels the construction of criterion-referenced tests (Popham, 1980; Hambleton, 1982c). Tables 3.2.1 and 3.2.2 provide information on the distribution of item content across six content categories for four test booklets (two booklets at the 9 and 13 year old levels). Math calculations, story problems, and geometry appear to be the most frequently occurring types of test items.

Table 3.2.1

Content Classification Summary of NAEP
Math Booklet No. 1 and 2 Test Items
(9 Year Olds, 1977-78)

	Booklet 1	Booklet 2
Story Problems		
Money	1	3
General	5	2
Logic, Probability, Permutation and Combination	4	7
Total	10	12
Geometry		
Story	0	0
Definition/Operations	9	9
Figure Interpretations, Manipulation	5	1
Total	14	10
Definition		
Total	1	16
Calculation		
General	15	25
Algebra	8	1
Total	23	26
Measurement		
English	3	1
Metric	3	4
Total	6	5
Graphs and Figures		
Total	5	6

Table 3.2.2

Content Classification Summary of NAEP
Math Booklet No. 1 and 2 Test Items
(13 Year Olds, 1977-78)

Booklet 1		Booklet 2	
Story Problems		**Story Problems**	
Money	3	Money	2
General	6	General	9
Logic, Probability, Permutation and Combination	5	Logic, Probability, Permutation and Combination	4
Total	14	Total	15
Geometry		**Geometry**	
Story	1	Story	1
Definition/Operations	9	Definition/Operations	7
Figure Interpretation, Manipulation	3	Figure Interpretation, Manipulation	2
Total	13	Total	10
Definition		**Definition**	
Total	9	Total	7
Calculation		**Calculation**	
General	14	General	17
Algebra	1	Algebra	5
Total	15	Total	22
Measurement		**Measurement**	
English	3	English	1
Metric	2	Metric	0
Total	5	Total	1
Graphs and Figures		**Graphs and Figures**	
Total	1	Total	7

About 50% of the test items in the 1977-78 assessment were included on the previous NAEP mathematics assessment in 1971-72. In addition, on the 1977-78 assessment some test items were included in the mathematics test booklets at all three age levels. While in our research investigation "linking" test items across age levels to a common scale was of no interest, such a task could have been accomplished with the aid of these common test items (Lord, 1980; Wright & Stone, 1979). However, at a given age level, there were no common test items. Had we been interested in "linking" test items at a given age level to a common scale, the task could have been achieved easily because of the plausible assumption that test forms were administered to equivalent ability groups (Lord, 1980; Hambleton & Swaminathan, 1982).

Test items in the NAEP mathematics assessment were of two types: multiple-choice, and open-ended. Tables 3.2.3, 3.2.4, 3.2.5, and 3.2.6 provide information on the item formats and content categories of test items in NAEP math booklets 1 and 2 for 9 and 13 year olds. Among the multiple-choice test items it was also interesting to note that the number of answer choices varied. Information reported in the four tables provided the basis for several important analyses described in section 3.6.

3.3 Description of Data

Six NAEP mathematics test booklets from the 1977-78 assessment were selected for analysis:

9 Year Olds

Booklet No. 1, 65 test items
Booklet No. 2, 75 test items
Booklet No. 3, 68 test items

Table 3.2.3

Format and Content Classification of NAEP
Math Booklet No. 1 Test Items
(9 Year Olds, 1977-78)

Item No.	Answer Format[1]	Category
1/102A	MC	Definition
2/102B	MC (6 options)	Definition
3/103A	MC	Story problem - money
4/104A	MC (6 options)	Geometry - definition
5/104B	MC (6 options)	Geometry - definition
6/105A	MC	Geometry - figure manipulation, interpretation
7/106A	OE	Geometry - operations
8/106B	OE	Geometry - operations
9/106C	MC	Geometry - operations
10/107A	MC (6 options)	Measurement - English
11/108A	OE	Calculation
12/108B	OE	Calculation
13/108C	OE	Calculation
14/108D	OE	Calculation
15/108E	OE	Calculation
16/108F	OE	Calculation
17/109A	MC	Story problem - logic
18/110A	OE	Story problem - general
19/111A	MC	Geometry - definition
20/112A	OE	Calculation
21/112B	OE	Calculation
22/113A	MC	Measurement - English
23/114A	MC (6 options)	Story problem - general
24/115A	OE	Calculation - algebra
25/115B	OE	Calculation - algebra
26/115C	OE	Calculation - algebra
27/115D	OE	Calculation - algebra
28/115E	OE	Calculation - algebra
29/115F	OE	Calculation - algebra
30/115G	OE	Calculation - algebra

---

[1] MC Items have 5 answer choices (including "I don't know") unless otherwise noted.

Table 3.2.3 (continued)

Item No.	Answer Format	Category
31/116A	MC (6 options)	Graphs and Figures
32/117A	MC	Definition
33/117B	MC	Definition
34/118A	MC (4 options)	Measurement - metric
35/119A	MC (6 options)	Graphs and Figures
36/120A	OE	Calculation
37/120B	OE	Calculation
38/121A	MC (10 options)	Definition
39/122A	MC (6 options)	Story problem -general
40/123A	MC	Calculation
41/124A	OE	Story problem - general
42/125A	OE	Calculation
43/125B	OE	Calculation
44/125C	OE	Calculation
45/126A	OE	Measurement - metric
46/127A	MC (4 options)	Calculation - algebra
47/128A	MC (4 options)	Measurement - metric
48/129A	MC	Graphs and Figures
49/129B	MC	Graphs and Figures
50/130A	MC (4 options)	Story problem - logic
51/130B	MC (4 options)	Story problem - logic
52/131A	MC (7 options)	Geometry - figure manipulation, interpretation
53/131B	MC (7 options)	Geometry - figure manipulation, interpretation
54/131C	MC (7 options)	Geometry - figure manipulation, interpretation
55/132A	OE	Graphs and Figures
56/133A	OE	Story problem - general
57/134A	MC (6 options)	Geometry - definition
58/134B	MC (6 options)	Geometry - definition
59/134C	MC (6 options)	Geometry - definition
60/135A	OE	Story problem - probability

Table 3.2.3 (continued)

Item No.	Answer Format	Category
61/136A	OE	Measurement - English
62/137A	OE	Definition
63/138A	OE	Calculation
64/139A	MC	Geometry - figure manipulation, interpretation
65/140A	MC	Definition

Table 3.2.4

Format and Content Classification of NAEP Math Booklet No. 2 Test Items (9 Year Olds, 1977-78)

Item No.	Answer Format[1]	Category
1/202A	MC	Definition
2/202B	MC	Definition
3/203A	OE	Calculation
4/203B	OE	Calculation
5/203C	OE	Calculation
6/203D	OE	Calculation
7/203E	OE	Calculation
8/203F	OE	Calculation
9/204A	OE	Calculation
10/204B	OE	Calculation
11/204C	OE	Calculation
12/204D	OE	Calculation
13/205A	MC (6 options)	Geometry - operations
14/206A	MC (6 options)	Story problem - money
15/207A	MC	Graphs and Figures
16/207B	MC	Graphs and Figures
17/208A	OE	Calculation
18/208B	OE	Calculation
19/208C	OE	Calculation
20/209A	MC	Story problem - combinations
21/210A	MC (8 options)	Graphs and Figures
22/210B	MC (6 options)	Graphs and Figures
23/210C	MC (9 options)	Graphs and Figures
24/211A	MC (4 options)	Definition
25/211B	MC (4 options)	Definition
26/211C	MC (4 options)	Definition
27/211D	MC (4 options)	Definition
28/211E	MC (4 options)	Definition
29/212A	MC (4 options)	Measurement - metric
30/212B	MC	Measurement - metric

[1] MC Items have 5 answer choices (including "I don't know") unless otherwise noted.

Table 3.2.4 (continued)

Item No.	Answer Format	Category
31/213A	OE	Calculation - algebra
32/214A	OE	Story problem - logic
33/215A	OE	Definition
34/215B	OE	Definition
35/215C	OE	Definition
36/216A	MC (6 options)	Geometry - definition
37/216B	MC (6 options)	Geometry - definition
38/216C	MC (6 options)	Geometry - definition
39/217A	MC	Story problem - money
40/218A	OE	Calculation
41/218B	OE	Calculation
42/218C	OE	Calculation
43/218D	OE	Calculation
44/218E	OE	Calculation
45/218F	OE	Calculation
46/219A	MC	Geometry - operations
47/220A	OE	Calculation
48/220B	OE	Calculation
49/220C	OE	Calculation
50/221A	MC	Geometry - definition
51/222A	MC	Measurement - metric
52/223A	MC	Definition
53/224A	MC	Definition
54/224B	MC	Definition
55/225A	MC	Story problem - logic
56/225B	MC	Story problem - logic
57/225C	MC	Story problem - logic
58/226A	MC	Story problem - general
59/226B	MC	Story problem - general
60/227A	MC	Calculation

Table 3.2.4 (continued)

Item No.	Answer Format	Category
61/228A	MC	Geometry - definition
62/228B	MC	Geometry - definition
63/229A	MC	Definition
64/229B	MC	Definition
65/229C	MC	Definition
66/230A	OE	Calculation
67/231A	OE	Story problem - money
68/232A	OE	Geometry - operations
69/233A	MC	Story problem - logic
70/234A	OE	Story problem - probability
71/235A	OE	Geometry - figure manipulation, interpretation
72/236A	OE	Calculation
73/237A	OE	Measurement - English
74/238A	OE	Graphs and Figures
75/239A	MC	Measurement - metric

Table 3.2.5

Format and Content Classification of NAEP Math Booklet No. 1 Test Items (13 Year Olds, 1977-78)

Item No.	Answer Format[1]	Category
1/102A	OE	Story problem - money
2/103A	MC	Definitions
3/103B	MC	Definitions
4/104A	OE	Measurement - English
5/105A	MC	Calculation
6/106A	MC	Geometry - definition, operations
7/106B	MC	Geometry - definition, operations
8/106C	MC	Geometry - definition, operations
9/107A	MC	Story problem - logic
10/108A	MC	Measurement - metric
11/109A	OE	Calculation - subtraction
12/109B	OE	Calculation - subtraction
13/109C	OE	Caluclation - subtraction
14/109D	OE	Calculation - subtraction
15/109E	OE	Calculation - subtraction
16/109F	OE	Calculation - subtraction
17/110A	MC (4 options)	Measurement - metric
18/111A	OE	Story problem - general
19/111B	OE	Calculation
20/112A	OE	Calculation
21/112B	OE	Calculation
22/113A	MC (10 options)	Definition
23/114A	MC	Definition
24/114B	MC	Definition
25/115A	OE	Story problem - money
26/116A	MC	Geometry - definitions, operations
27/116B	MC	Geometry - definitions, operations
28/117A	OE	Geometry - definitions
29/118A	OE	Measurement - English
30/119A	MC (7 options)	Story problems - general

---

[1] MC items have 5 answer choices (including "I don't know") unless otherwise noted.

Table 3.2.5 (continued)

Item No.	Answer Format	Category
31/120A	MC	Geometry - figure manipulation, interpretation
32/120B	MC	Story problem - general
33/121A	MC (6 options)	Story problem - general
34/122A	MC	Geometry - definitions
35/122B	MC	Geometry - definitions
36/123A	MC	Story problem - money
37/124A	MC (6 options)	Geometry - story problem
38/125A	MC	Definitions
39/126A	MC	Definitions
40/127A	MC	Story problem - combinations
41/128A	MC	Definitions
42/129A	MC (6 options)	Geometry - definitions, operations
43/130A	MC	Geometry - figure manipulation
44/131A	OE	Calculation
45/131B	OE	Calculation
46/132A	MC	Story problem - general
47/133A	MC	Geometry - story problem
48/134A	MC (6 options)	Definitions
49/135A	OE	Calculations - algebra
50/136A	MC	Story problem - general
51/137A	MC (6 options)	Story problem - probability
52/137B	MC (6 options)	Story problem - probability
53/138A	MC (6 options)	Geometry - figure manipulation
54/139A	OE	Calculation
55/140A	OE	Graphs and figures
56/141A	MC	Story problem - logic
57/142A	OE	Measurement - English
58/143A	OE	Calculation

Table 3.2.6

Format and Content Classification of NAEP
Math Booklet No. 2 Test Items
(13 Year Olds, 1977-78)

Item No.	Answer Format[1]	Category
1/202A	OE	Calculation - algebra
2/203A	OE	Calculation
3/204A	OE	Calculation
4/205A	MC	Story problem - logic
5/206A	MC	Definitions
6/207A	OE	Graphs and Figures
7/208A	OE	Measurement - English
8/209A	OE	Story problem - general
9/210A	OE	Calculation
10/210B	OE	Calculation
11/210C	OE	Calculation
12/210D	OE	Calculation
13/211A	MC (6 options)	Geometry - definitions
14/212A	MC	Calculation - algebra
15/213A	MC (6 options)	Geometry - story problem
16/214A	OE	Calculation
17/214B	OE	Calculation
18/214C	OE	Calculation
19/214D	OE	Calculation
20/214E	OE	Calculation
21/214F	OE	Calculation
22/215A	MC (6 options)	Geometry - definitions
23/216A	OE	Calculation
24/216B	OE	Calculation
25/216C	OE	Calculation
26/217A	MC	Geometry - definition
27/217B	MC	Geometry - definition
28/218A	OE	Story problem - general
29/219A	MC	Story problem - money
30/220A	OE	Story problem - probability

---

[1] MC items have 5 answer choices (including "I don't know") unless otherwise noted.

Table 3.2.6 (continued)

Item No.	Answer Format	Category
31/221A	MC	Definition
32/222A	MC (4 options)	Definition
33/222B	MC (4 options)	Definition
34/223A	MC (6 options)	Story problem - general
35/224A	OE	Story problem - money
36/225A	MC (6 options)	Graphs and figures
37/225B	MC (7 options)	Graphs and figures
38/225C	MC (6 options)	Graphs and figures
39/226A	OE	Calculation - algebra
40/227A	MC	Story problem - general
41/228A	OE	Calculation - algebra
42/228B	OE	Calculation - algebra
43/229A	MC (4 options)	Story problem - general
44/230A	MC	Geometry - figure manipulation, interpretation
45/231A	MC (6 options)	Story problem - permutation and combination
46/232A	MC	Story problem - general
47/232B	MC	Story problem - general
48/233A	OE	Definition
49/233B	OE	Definition
50/233C	OE	Definition
51/234A	MC (6 options)	Geometry - definitions
52/234B	MC (6 options)	Geometry - definitions
53/235A	OE	Story problem - general
54/236A	MC	Geometry - figure manipulation, interpretation
55/237A	MC (6 options)	Geometry - definitions, operations
56/238A	OE	Story problem - general
57/239A	MC (6 options)	Story problem - probability
58/240A	MC (6 options)	Graphs and figures
59/240B	MC (6 options)	Graphs and figures
60/240C	MC (6 options)	Graphs and figures
61/241A	OE	Calculation - algebra
62/241B	OE	Calculation - algebra

13 Year Olds

Booklet No. 1, 58 test items
Booklet No. 2, 62 test items
Booklet No. 3, 73 test items

In some of the computer printouts which follow the six booklets above are designated 109, 209, 309, 113, 213, 313, respectively. There was no particular pattern to our choice of data sets for the various analyses. For some analyses all six data sets were used, for others, only one or two were used.

Tables 3.3.1 and 3.3.2 contain the one- and three-parameter logistic model parameter estimates for items in the six NAEP math booklets mentioned above. Between 2400 and 2500 examinees were used in item parameter estimation which was carried out with the aid of LOGIST (Wingersky, 1982; Wingersky, Barton, & Lord, 1982[1]).

---

[1] The most recent references to LOGIST are given but the 1976 version of the computer program was used in our analyses.

Table 3.3.1

NAEP Math Item Response Model Parameter Estimates
(9 Year Olds, 1977-78)

Test Item	Booklet No. 1				Booklet No. 2				Booklet No. 3			
	1-p	3-p			1-p	3-p			1-p	3-p		
	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$
1	-.22	.20	1.15	.19	-1.39	-2.86	.24	.09	.44	2.67	.08	.01
2	.17	.30	1.20	.09	-1.40	-2.82	.25	.09	-.31	-1.76	.09	.01
3	-.22	.15	1.20	.17	-2.63	-2.36	.77	.09	.15	.24	.27	.01
4	-2.55	-4.01	.37	.06	-2.13	-1.64	.99	.09	-.14	-.54	.15	.01
5	-2.33	-3.39	.40	.06	-2.21	-1.86	.85	.09	-1.58	-4.96	.17	.01
6	-.93	-1.77	.27	.06	-1.40	-1.07	.92	.09	-1.23	-1.24	.64	.01
7	2.18	1.91	1.56	.07	1.99	-1.49	1.05	.09	3.62	3.22	.76	.01
8	.82	.86	.70	.03	-1.68	-1.26	1.00	.09	-1.49	-2.31	.37	.01
9	.21	.38	.51	.05	-.42	-.10	1.58	.12	.93	48.36	.01	.01
10	.53	.58	1.13	.08	-.48	-.20	1.42	.09	-1.63	-1.97	.50	.01
11	-2.32	-1.62	1.42	.06	.01	.09	1.48	.05	.04	.69	1.22	.27
12	-1.81	-1.30	1.26	.06	-.01	.08	1.26	.05	-1.94	-5.11	.21	.01
13	-2.17	-1.48	1.55	.06	2.63	2.18	1.07	.02	.83	.70	.82	.01
14	-1.13	-.79	1.24	.06	.60	.71	.95	.09	1.65	1.52	.71	.01
15	-1.62	-1.09	1.49	.06	-1.13	-1.00	.66	.09	1.15	.81	1.14	.00
16	-1.20	-.81	1.39	.06	-.32	.06	1.03	.17	6.31	4.83	.88	.00
17	.19	.51	.32	.06	-1.18	-.93	.82	.09	.31	.29	.89	.02
18	-1.64	-1.63	.65	.06	-1.14	-.88	.83	.09	-.02	-.05	.53	.01
19	-1.90	-2.23	.52	.06	-.26	-.07	.99	.09	-1.50	-1.28	.88	.01
20	-.60	-.47	.79	.06	1.99	1.64	1.07	.03	-2.00	-1.69	.89	.01
21	.48	.37	1.19	.01	-.60	-.33	1.06	.09	-2.73	-4.04	.40	.01
22	1.55	1.25	1.46	.06	-.32	-.10	1.19	.09	-1.58	-1.59	.66	.01
23	-.14	.32	1.39	.21	.52	.49	1.25	.05	-1.64	-1.64	.67	.01
24	-1.69	-1.22	1.20	.06	-3.50	-2.87	.96	.09	-1.27	-1.62	.47	.01
25	.49	.43	1.04	.03	-3.48	-3.12	.81	.09	-.28	-.21	1.02	.01
26	.06	.10	1.35	.03	-3.62	-2.93	1.00	.09	-1.91	-1.51	1.05	.01
27	-.82	-.56	1.06	.06	-3.03	-3.07	.65	.09	-1.75	-1.34	1.11	.01
28	.68	.60	.86	.02	-2.67	-2.71	.64	.09	-1.17	-.85	1.18	.01
29	-1.35	-1.07	.93	.06	1.58	2.45	1.10	.16	-.81	-.62	1.05	.01
30	-.83	-.53	1.33	.06	1.55	2.38	.56	.09	-1.41	-1.06	1.14	.01

Table 3.3.1 (continued)

NAEP Math Item Response Model Parameter Estimates

(9 Year Olds, 1977-78)

Test Item	Booklet No. 1				Booklet No. 2				Booklet No. 3			
	1-p	3-p			1-p	3-p			1-p	3-p		
	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$
31	-.89	-.66	.98	.06	1.20	1.06	.80	.00	-1.00	-.75	1.10	.01
32	3.59	2.51	2.00	.02	1.10	.88	.94	.00	4.23	3.57	.85	.01
33	1.61	8.93	.12	.06	.03	.18	1.15	.09	.91	1.41	.77	.15
34	-.64	-.63	.56	.06	-.10	.08	1.13	.09	-2.06	-1.97	.73	.01
35	-1.51	-1.41	.71	.06	.13	.27	1.10	.09	-.26	-.97	.15	.01
36	-.68	-.51	.89	.06	.50	1.38	.87	.26	-1.11	-.94	.87	.01
37	2.47	1.88	1.00	.00	-1.87	-1.74	.69	.09	1.68	1.45	.79	.01
38	-2.99	-3.93	.47	.06	1.60	1.68	.91	.07	-1.47	-1.71	.54	.01
39	1.53	2.52	.67	.13	1.21	1.86	1.79	.21	-3.45	-4.02	.58	.01
40	1.77	2.40	1.00	.13	-.61	-.35	1.09	.09	-4.21	-5.24	.60	.01
41	1.55	1.07	1.33	.00	.46	.35	1.17	.01	5.57	3.66	1.11	.00
42	1.77	1.58	.77	.00	-1.07	-.65	1.35	.09	5.48	3.76	1.04	.00
43	4.02	2.53	1.40	.00	-.25	-.15	1.40	.01	5.66	3.49	1.23	.00
44	4.90	2.80	1.62	.00	-1.20	-.74	1.42	.09	5.76	3.78	1.10	.00
45	1.03	2.56	.27	.06	.18	.15	1.21	.01	5.87	3.32	2.00	.01
46	.69	.96	1.18	.16	1.58	1.77	1.15	.11	2.08	1.88	.73	.01
47	-.18	-.15	.37	.06	-2.37	-3.15	.44	.09	1.51	1.22	1.09	.04
48	-1.74	-1.51	.82	.06	-1.23	-.93	.88	.09	-2.62	-2.48	.76	.01
49	1.02	1.14	1.38	.13	.00	.16	.85	.09	-2.58	-2.15	.98	.01
50	-.72	-1.48	.24	.06	.39	.58	.75	.09	-2.41	-1.95	1.05	.01
51	1.23	3.82	.21	.06	-1.49	-3.33	.23	.09	-2.05	-1.75	.90	.01
52	.71	3.32	.14	.06	-1.04	-.63	1.27	.09	-2.21	-1.96	.85	.01
53	2.49	2.49	1.51	.07	-.01	.39	1.44	.20	-2.22	-1.91	.90	.01
54	2.55	2.39	1.17	.06	.28	.73	.76	.15	-.16	-.27	.39	.01
55	5.06	3.11	1.44	.00	1.20	2.30	.60	.15	1.98	3.37	.69	.12
56	2.14	1.60	1.05	.00	-.17	.02	.45	.09	-1.43	-1.27	.82	.01
57	.12	.48	.82	.15	.60	1.21	.65	.16	-.86	-.84	.69	.01
58	1.32	1.33	1.08	.09	-.74	-.56	.73	.09	-.10	-.25	.28	.01
59	1.06	1.20	1.62	.15	.01	.29	.94	.13	1.36	2.02	1.01	.17
60	1.51	1.55	.63	.01	1.33	1.34	.92	.06	-.96	-1.05	.59	.01

Table 3.3.1 (continued)
NAEP Math Item Response Model Parameter Estimates
(9 Year Olds, 1977-78)

Test Item	Booklet No. 1				Booklet No. 2				Booklet No. 3			
	1-p	3-p			1-p	3-p			1-p	3-p		
	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$
61	2.41	1.88	1.13	.02	-2.18	-4.00	.09	.09	.73	.78	1.08	.09
62	-1.95	-1.94	.66	.06	.30	.83	.35	.09	-.07	.18	1.42	.12
63	.09	.17	.94	.06	1.49	2.43	1.26	.18	.43	.68	.33	.01
64	.70	.99	.71	.10	2.01	2.49	1.79	.12	1.31	1.27	.94	.06
65	1.12	3.31	.22	.06	1.21	4.72	.19	.09	-1.41	-1.39	.68	.01
66					2.12	2.14	.75	.03	-.51	-1.03	.28	.01
67					1.14	1.45	.68	.09	.70	.56	.91	.01
68					4.82	3.15	1.23	.00	-.66	-.84	.49	.01
69					.06	.35	.46	.09				
70					2.61	4.84	.38	.03				
71					2.88	3.16	.65	.02				
72					3.28	2.38	1.14	.00				
73					.74	1.15	.57	.09				
74					.49	.64	.80	.09				
75					.73	1.51	.39	.09				

Table 3.3.2
NAEP Math Item Response Model Parameter Estimates
(13 Year Olds, 1977-78)

Test Item	Booklet No. 1				Booklet No. 2				Booklet No. 3			
	1-p	3-p			1-p	3-p			1-p	3-p		
	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$
1	-1.92	-1.43	1.00	.11	-.32	-.26	.71	.04	-1.51	-.86	1.52	.11
2	-3.71	-2.40	.77	.11	.10	.28	1.10	.10	-1.68	-1.00	1.48	.11
3	-2.19	-2.87	.77	.11	-.67	-.77	.54	.04	-1.13	-1.27	.43	.11
4	-.09	.03	1.72	.04	-.86	-1.07	.49	.04	-.39	-.14	.87	.11
5	-.67	-.34	1.15	.11	-.30	-.02	1.13	.13	.39	.39	1.21	.06
6	.66	1.05	1.12	.17	1.56	1.58	.72	.00	-2.63	-3.12	.48	.11
7	.45	.94	1.19	.21	-.03	.05	.78	.04	-.79	-.47	.90	.11
8	-.94	-1.78	.24	.11	-3.13	-3.51	.63	.04	-2.06	-1.58	.89	.11
9	.95	1.39	.67	.11	-1.78	-1.25	1.66	.04	1.51	2.21	1.00	.16
10	-1.60	-1.12	1.06	.11	-1.72	-1.21	1.65	.04	1.29	2.59	1.21	.22
11	-3.22	-2.27	1.31	.11	-1.61	-1.15	1.58	.04	-2.07	-1.73	.77	.11
12	-2.88	-2.09	1.20	.11	-1.40	-1.01	1.45	.04	.84	.60	1.10	.00
13	-2.72	-1.94	1.23	.11	-2.51	-2.35	.83	.04	.15	.16	1.15	.03
14	-2.65	-2.00	1.05	.11	.33	.52	.54	.04	-2.29	-4.06	.29	.11
15	-2.28	-1.83	.91	.11	.91	.90	.99	.04	-.04	.45	1.16	.21
16	-2.20	-1.96	.75	.11	-2.16	-2.06	.79	.04	.19	.61	1.15	.20
17	-1.08	-1.16	.48	.11	-1.78	-1.61	.86	.04	.31	.70	1.32	.20
18	2.02	1.86	.78	.01	-1.89	-1.71	.87	.04	1.06	2.19	2.00	.26
19	-.54	-.25	1.23	.11	-3.03	-3.75	.55	.04	-1.08	-.66	1.12	.11
20	-.42	-.20	1.92	.04	-3.03	-2.84	.84	.04	.69	.86	1.24	.14
21	-.25	-.12	1.71	.02	-2.89	-3.25	.62	.04	-.36	-.10	.98	.11
22	-2.84	-3.41	.50	.11	2.70	3.53	2.00	.06	-2.84	-2.61	.71	.11
23	.64	1.27	1.51	.25	.53	.44	1.18	.01	-3.00	-2.48	.86	.11
24	.81	1.13	1.12	.16	.22	.23	1.18	.03	-2.94	-2.40	.88	.11
25	-.07	.14	.94	.11	-.48	-.40	.78	.04	-3.30	-2.68	.92	.11
26	-1.10	-1.39	.39	.11	-2.02	-6.17	.19	.04	-3.37	-2.67	.97	.11
27	2.34	3.04	.96	.02	-.19	-.23	.32	.04	-3.12	-2.43	.99	.11
28	1.43	2.75	.42	.15	1.19	1.24	1.39	.10	.14	.47	.84	.14
29	1.65	1.46	.87	.02	-2.39	-2.57	.66	.04	1.62	2.02	.82	.11
30	.21	.48	1.18	.15	1.67	1.71	.71	.00	.32	1.16	.30	.11

Table 3.3.2 (continued)
NAEP Math Item Response Model Parameter Estimates
(13 Year Olds, 1977-78)

Test Item	Booklet No. 1				Booklet No. 2				Booklet No. 3			
	1-p	3-p			1-p	3-p			1-p	3-p		
	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$
31	-1.18	-.88	.84	.11	.87	.78	1.10	.03	1.90	2.80	2.00	.16
32	-.34	.16	1.46	.21	-.75	-1.03	.42	.04	-1.28	-1.03	.71	.11
33	.36	.66	.68	.11	-2.93	-2.43	1.06	.04	.40	.66	1.02	.14
34	-3.25	-4.13	.47	.11	-1.90	-1.81	.78	.04	1.49	11.26	.10	.11
35	-.75	-.73	.46	.11	1.29	1.30	.72	.00	-.02	.20	.97	.11
36	1.44	3.75	2.00	.21	-.39	-.31	.79	.04	1.16	87.54	.01	.11
37	.63	.97	.69	.11	-.75	-.44	1.30	.10	-.28	-.01	.71	.11
38	-1.41	-.88	1.37	.11	.68	.76	1.07	.07	-.16	.32	.13	.11
39	-.93	-1.24	.35	.11	.01	.08	.84	.04	2.13	2.45	.79	.07
40	-.71	-.41	1.05	.11	.20	.40	1.07	.10	-.84	-.46	1.11	.11
41	1.12	1.00	1.11	.05	-2.01	-1.81	.88	.04	.12	.52	.45	.11
42	-.89	-.62	.79	.11	-1.33	-1.13	.96	.04	1.68	3.62	2.00	.19
43	-1.39	-1.16	.73	.11	-1.32	-1.39	.65	.04	2.25	1.54	1.19	.00
44	-.82	-.65	.69	.11	-1.04	-1.86	.32	.04	-.86	-1.52	.23	.11
45	.23	.42	.85	.11	-.70	-.78	.56	.04	.88	1.02	1.17	.13
46	.74	.74	1.28	.08	-1.51	-1.37	.85	.04	3.63	3.15	1.18	.03
47	2.24	3.06	1.54	.11	-1.45	-1.17	1.07	.04	.09	.48	1.00	.18
48	1.92	1.86	.89	.04	1.07	.94	1.02	.16	.70	1.04	1.95	.20
49	.04	.24	.97	.11	1.63	1.27	1.20	.00	.77	.96	1.79	.17
50	-1.82	-1.57	.76	.11	-.57	-.42	1.02	.04	.53	.97	2.00	.24
51	1.67	2.24	1.03	.13	.78	4.44	.12	.04	2.43	1.95	.90	.01
52	-.46	-.71	.20	.11	1.62	2.65	.58	.08	.48	.78	.98	.15
53	-1.68	-2.04	.45	.11	.48	.39	1.18	.00	.06	.14	1.03	.05
54	-1.10	-.86	.78	.11	-1.87	-2.21	.57	.04	.50	.36	1.10	.00
55	1.24	.95	1.13	.01	.12	.22	.66	.04	-.13	.03	1.09	.07
56	-1.05	-.85	.71	.11	.02	.07	1.44	.03	2.41	2.58	.79	.05
57	.90	.66	1.24	.00	1.95	2.57	2.00	.12	1.64	1.96	1.15	.13
58	-1.19	-.93	.79	.11	1.64	1.43	1.29	.04	-2.61	-1.84	1.13	.11
59					1.51	1.46	1.63	.09	-1.66	-1.30	.81	.11
60					-1.16	-1.28	.60	.04	-1.10	-1.01	.55	.11

Table 3.3.2 (continued)
NAEP Math Item Response Model Parameter Estimates
(13 Year Olds, 1977-78)

Test Item	Booklet No. 1				Booklet No. 2				Booklet No. 3			
	1-p	3-p			1-p	3-p			1-p	3-p		
	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{b}$	$\hat{b}$	$\hat{a}$	$\hat{c}$
61					-.62	-.62	.64	.04	.80	.56	1.33	.01
62					.95	.78	1.07	.00	.36	.64	.67	.11
63									.90	.61	1.40	.01
64									-.59	-.42	.54	.11
65									2.45	1.80	1.13	.02
66									-.61	-.38	.67	.11
67									2.07	1.64	.92	.01
68									.25	.22	1.01	.02
69									.07	.21	1.12	.11
70									3.21	2.31	1.01	.00
71									.17	.15	1.14	.02
72									.22	.17	1.05	.00
73									-.97	-.82	.60	.11

3.4 Checking Model Assumptions

Checking on two model assumptions, unidimensionality and equal item discrimination indices, with respect to the NAEP math booklets, was carried out. The results will be presented next. It was not necessary to check the level of test speededness because test items were administered one at a time to examinees and they were given sufficient time on each one to provide answers.

Unidimensionality

Checks on the unidimensionality of the math booklets were carried out with NAEP Math Booklet No. 1 for 13 year olds. A study of the eigenvalues was carried out with Black and White samples of 330 examinees. The samples were drawn at random from the available pool of examinees taking the math booklet. The largest eigenvalues for each sample are presented in order in Table 3.4.1. Comparable results from the Black and White samples were obtained: About 15% of the total variance was accounted for by the first factor or component and the ratio of the first to the second eigenvalue was (approximately) 2.8 (2.7 in the Black sample and 2.9 in the White sample)[1]. These statistics do not meet Reckase's (1979) minimal criteria for unidimensionality. However, since his criteria are arbitrary and other goodness of fit evidence would be available, the decision made was to move on to other types of analyses.

Equal Item Discrimination Indices

Table 3.4.2 provides item difficulty and discrimination (biserial correlations) information for NAEP Math Booklet No. 1 for 13 Year Olds. The information is reported for six groups: Two Black and White samples

---

[1] Somewhat better results were obtained from an analysis of the total sample (N=2422). About 17.6% of the variance was accounted for by the first factor and the ratio of the first to the second eigenvalue was 3.6.

Table 3.4.1

Listing of the Largest Eigenvalues
for NAEP Math Booklet No. 1
(13 year olds, 1977-78)

Eigenvalue	Black Sample (N=330)	White Sample (N=330)
1	8.4	8.8
2	3.1	3.0
3	1.9	2.1
4	1.8	1.8
5	1.7	1.7
6	1.5	1.6
7	1.5	1.5
8	1.5	1.5
9	1.4	1.4
10	1.3	1.3
11	1.3	1.3
12	1.3	1.3
13	1.2	1.2
14	1.2	1.2
15	1.2	1.1
16	1.1	1.1
17	1.1	1.1
18	1.1	1.1
19	1.0	1.0
20	1.0	1.0
% Variance	14.4%	15.2%

Table 3.4.2

Summary of Item Statistics for NAEP Math Booklet No. 1
(13 Year Olds, 1977-78)

Item	Item Difficulty Level						Item Discrimination Index					
	Group[1]						Group					
	Black 1	Black 2	Black	White 1	White 2	White	Black 1	Black 2	Black	White 1	White 2	White
1	.68	.68	.68	.88	.90	.89	.60	.63	.61	.79	.81	.80
2	.82	.84	.83	.93	.95	.94	.55	.41	.48	.38	.59	.46
3	.88	.87	.88	.96	.96	.96	.57	.53	.55	.25	.68	.48
4	.12	.13	.12	.61	.65	.63	.66	.88	.77	.72	.70	.71
5	.39	.42	.40	.75	.66	.71	.70	.65	.67	.68	.66	.67
6	.34	.22	.28	.41	.41	.41	.22	.26	.24	.59	.45	.52
7	.32	.27	.30	.44	.49	.47	.29	.28	.29	.61	.43	.51
8	.63	.66	.65	.71	.74	.72	.26	.16	.21	.04	.41	.22
9	.19	.15	.17	.36	.30	.33	.22	.29	.25	.44	.53	.48
10	.54	.50	.52	.89	.84	.86	.66	.58	.62	.53	.74	.65
11	.85	.87	.86	.98	.98	.98	.78	.56	.68	.75	.68	.72
12	.84	.83	.84	.98	.95	.96	.80	.69	.74	.52	.72	.65
13	.83	.78	.81	.95	.97	.96	.72	.60	.65	.66	.88	.75
14	.82	.78	.80	.96	.96	.96	.70	.51	.60	.65	.86	.75
15	.78	.74	.76	.96	.93	.95	.51	.59	.55	.61	.71	.67
16	.79	.77	.78	.90	.93	.91	.65	.59	.62	.53	.56	.53
17	.50	.53	.51	.73	.78	.76	.32	.49	.40	.41	.32	.37
18	.06	.04	.05	.22	.16	.19	.76	.57	.68	.56	.36	.47
19	.33	.28	.31	.75	.70	.72	.77	.76	.77	.72	.69	.70
20	.24	.25	.25	.67	.72	.69	.89	.83	.86	.82	.72	.77
21	.25	.25	.25	.62	.67	.65	.81	.92	.86	.75	.69	.72
22	.82	.86	.84	.92	.96	.94	.52	.26	.40	.38	.84	.52
23	.36	.28	.32	.39	.39	.40	.29	.12	.21	.40	.40	.40
24	.22	.23	.23	.38	.33	.35	.29	.40	.35	.45	.48	.47
25	.31	.30	.30	.58	.61	.59	.70	.61	.66	.58	.70	.64

[1] Sample Sizes are as follows: Black 1 = Black 2 = White 1 = White 2 = 165; Black = White = 330.

Table 3.4.2 (continued)

Item	Item Difficulty Level						Item Discrimination Index					
	Group						Group					
	Black 1	Black 2	Black	White 1	White 2	White	Black 1	Black 2	Black	White 1	White 2	White
26	.61	.61	.61	.75	.73	.74	.18	.19	.18	.38	.42	.40
27	.02	.01	.02	.12	.10	.11	.13	.59	.23	.74	.40	.58
28	.21	.16	.19	.24	.18	.21	.17	.27	.22	.18	.22	.20
29	.05	.06	.05	.25	.19	.22	.45	.38	.41	.54	.56	.55
30	.19	.24	.22	.52	.53	.52	.39	.60	.50	.59	.49	.54
31	.55	.56	.55	.78	.81	.79	.55	.59	.57	.65	.66	.66
32	.28	.32	.30	.65	.62	.64	.42	.41	.41	.63	.49	.56
33	.24	.32	.28	.48	.47	.48	.51	.34	.41	.60	.56	.58
34	.85	.92	.88	.96	.97	.97	.31	.41	.35	.05	.57	.28
35	.51	.53	.52	.74	.68	.71	.22	.28	.25	.50	.43	.47
36	.29	.24	.27	.23	.15	.19	.14	.07	.11	.00	-.07	-.02
37	.15	.14	.14	.39	.42	.40	.28	.43	.35	.48	.36	.42
38	.48	.51	.50	.87	.86	.87	.74	.82	.77	.77	.87	.83
39	.56	.61	.58	.70	.72	.71	.28	.44	.35	.46	.34	.40
40	.33	.33	.33	.72	.76	.74	.68	.64	.66	.67	.68	.67
41	.12	.10	.11	.27	.31	.29	.58	.22	.41	.66	.46	.55
42	.47	.47	.47	.72	.74	.73	.40	.54	.47	.51	.61	.56
43	.46	.50	.48	.82	.79	.80	.50	.34	.42	.52	.63	.57
44	.44	.39	.41	.77	.76	.76	.68	.65	.66	.52	.36	.44
45	.21	.27	.24	.44	.48	.46	.66	.47	.55	.53	.70	.61
46	.16	.15	.15	.40	.38	.38	.42	.63	.52	.72	.51	.61
47	.13	.08	.11	.08	.12	.10	.19	.03	.12	.07	.05	.05
48	.05	.08	.06	.19	.15	.17	.20	.16	.17	.68	.23	.48
49	.28	.29	.28	.56	.52	.54	.62	.62	.62	.58	.64	.61
50	.64	.73	.68	.90	.87	.88	.57	.54	.55	.52	.85	.69

Table 3.4.2 (continued)

Item	Item Difficulty Level						Item Discrimination Index					
	Group						Group					
	Black 1	Black 2	Black	White 1	White 2	White	Black 1	Black 2	Black	White 1	White 2	White
51	.12	.13	.12	.22	.21	.22	.03	.16	.10	.26	.40	.33
52	.50	.49	.50	.65	.68	.67	.16	-.06	.04	.32	.28	.30
53	.70	.70	.70	.85	.83	.84	.33	.61	.47	.49	.39	.43
54	.55	.49	.52	.79	.72	.75	.67	.56	.61	.60	.49	.54
55	.06	.06	.06	.32	.30	.31	.74	.65	.70	.72	.58	.65
56	.48	.51	.49	.84	.71	.77	.51	.57	.54	.45	.58	.52
57	.08	.05	.06	.40	.42	.41	.87	.93	.89	.69	.62	.66
58	.48	.54	.51	.85	.79	.82	.51	.48	.49	.47	.76	.62

of 165 examinees, and combined Black and combined White samples of 330 examinees. The results for the four smaller samples are reported here although they were of more importance for an analysis reported later. The discrimination indices obtained with the Black and White samples reveal two things: The item discrimination indices for the two samples are comparable (average absolute difference = .11; Black values were higher 22 times; White values were higher for 35 items) although the White values tended to be a little higher; and more importantly, there is substantial variation among the item discrimination indices (.04 to .89 in the Black sample and -.02 to .83 in the White sample).

In a more complete analysis the following results were obtained:

Booklet	Sample Size	Test Length	Item Discrimination Indices[1] Mean	SD
Booklet No. 1, 9 Year Olds	2495	65	.565	.260
Booklet No. 2, 9 Year Olds	2463	75	.565	.260
Booklet No. 1, 13 Year Olds	2500	58	.585	.250
Booklet No. 2, 13 Year Olds	2433	62	.615	.252

The results above show clearly that the assumption of equal item discrimination indices is violated to a considerable degree. This finding is not surprising because item statistics play only a small part in NAEP mathematics test development. It would be reasonable, therefore, to expect a wider range of values than might be found on a standardized achievement or aptitude test where items with low discrimination indices are most likely deleted. Therefore, it is reasonable to suspect that the two- or three-parameter logistic models will provide a more adequate fit to the test results. This point is addressed in more detail in section 3.6.

---

[1] Correlations were transformed via Fisher's Z transformation prior to calculating the descriptive statistics. The mean is reported on the correlation scale. The standard deviation is reported on the Z_r scale.

3.5 Checking Model Features

When an item response model fits a test data set, at least to an adequate degree, two advantages or features are obtained: (1) item parameter estimates do not depend upon the samples of examinees drawn from the population of examinees for whom the test is designed (i.e., item parameter invariance) and (2) expected values of ability estimates do not depend upon the choice of test items. The extent to which the first feature was obtained with NAEP math data will be presented next.

Item Parameter Invariance

The invariance of item difficulty estimates for Whites and Blacks with the one-parameter model was investigated initially with Math Booklet No. 1 for 13 Year Olds. Three hundred and thirty Black examinees were located on the NAEP data tape. All these examinees were used in the analysis. An equal number of White students were selected at random from the same data tape. Next, the Black and the White student samples were divided at random into two halves so that four equal-sized (N=165) groups of students could be obtained. These groups were labelled "White 1," "White 2," "Black 1," and "Black 2." A one-parameter analysis was carried out with each group. The plots of "b" values in the two White and Black samples are shown in Figures 3.5.1 and 3.5.2. The plots show high relationships between the sets of b values ($r \doteq .98$). What variation there is in the plots is due to model-data misfit and examinee sampling errors. The plots provide a basis for investigating hypotheses

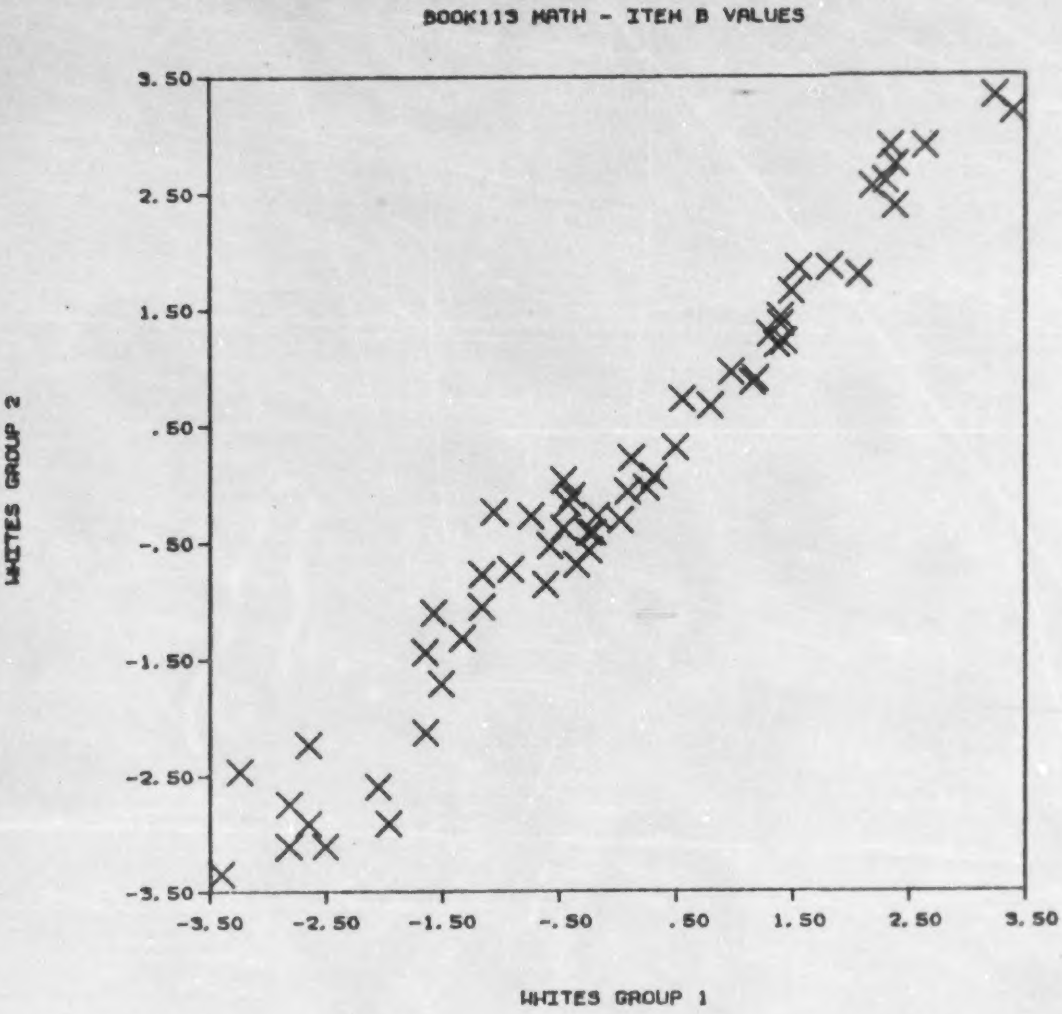


Figure 3.5.1 Plot of b values for the one-parameter model obtained from two equivalent white student samples (N=165).

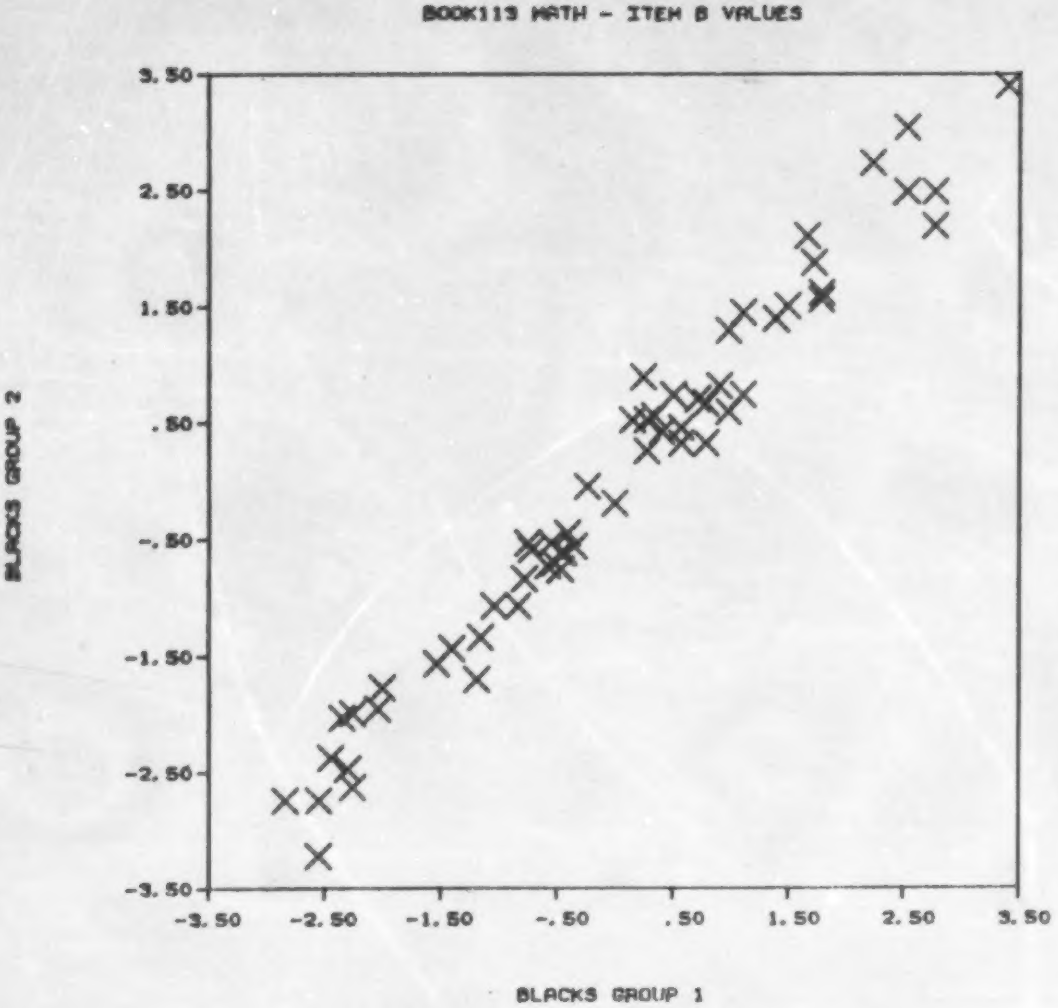


Figure 3.5.2 Plot of b values for the one-parameter model obtained from two equivalent black student samples (N=165).

concerning the invariance of item parameter estimates. If the feature of item invariance is present, similar plots should be obtained when the Black and White item parameter estimates are compared. Figure 3.5.3 reveals clearly that item difficulty estimates differ substantially in the first Black and White samples ($r \cong .74$). Figure 3.5.4 provides a replication of the Black-White comparison of item difficulty estimates. The plot of b values in Figure 3.5.4 is very similar to the plot in Figure 3.5.3 and both plots differ substantially from the baseline plots shown in Figure 3.5.1 and 3.5.2.

Figure 3.5.5 provides a plot of the differences in item difficulty estimates between the two White and the two Black samples ($r = .06$). The item parameter estimates obtained in each racial group should estimate the same item parameter value if the feature of item invariance is obtained (although the value may be different in the two racial groups). Therefore, the expected differences should be zero and the correlation of these differences across the set of test items in these two racial groups should also be zero. In fact, the correlation is very close to zero. If the feature of item invariance is present it should exist for any pairings of the data. Figure 3.5.6 shows that the correlation between b value differences in the first and second Black and White samples is not zero (in fact, $r = .72$!). Clearly, item difficulty estimates obtained with the one-parameter model are not invariant in the Black and White examinee samples.

The appropriate conclusion seems to be that item invariance across the two racial groups is not obtained. However, we stop short here of attributing the problem to race bias in the test items. There are at least two other plausible explanations: (1) the problem is due to a variable which is

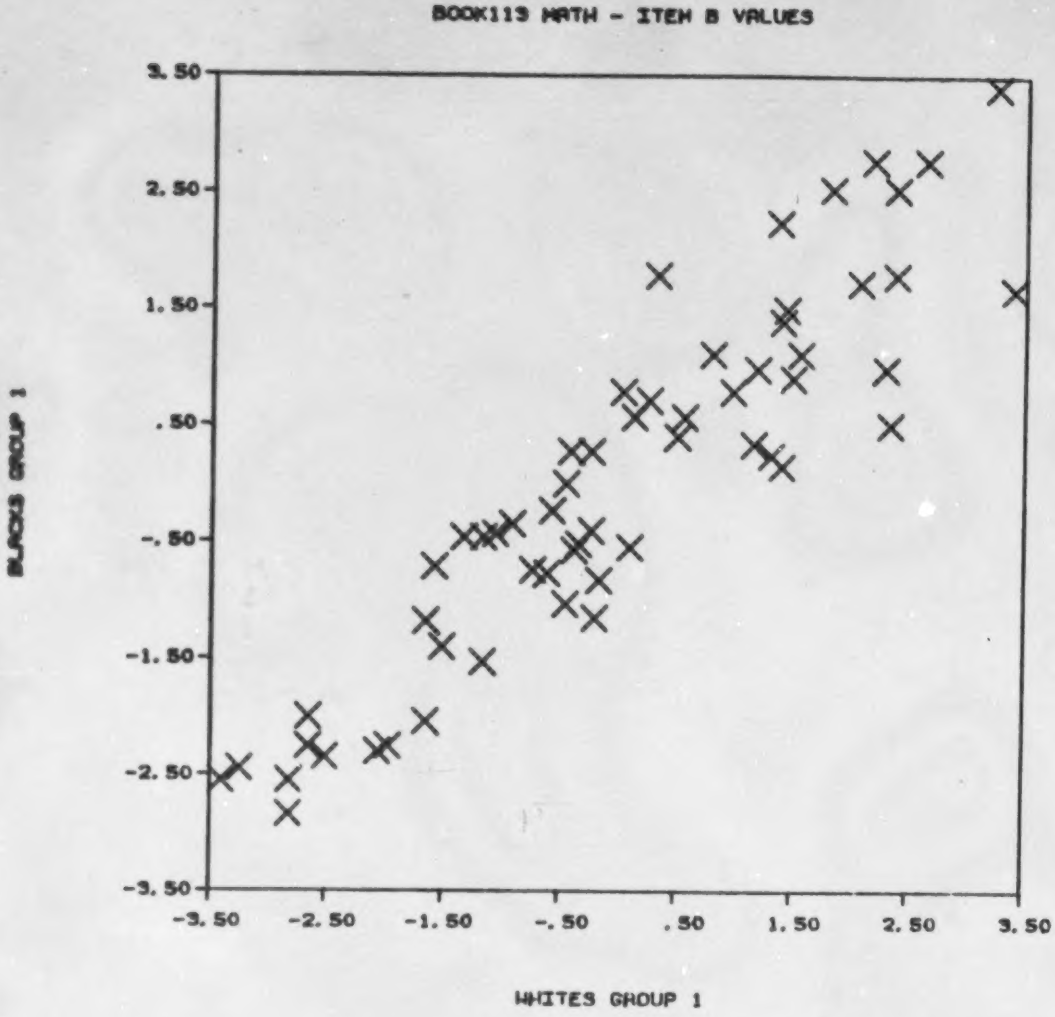


Figure 3.5.3 Plot of b values for the one-parameter model obtained from the first white and black student samples (N=165).

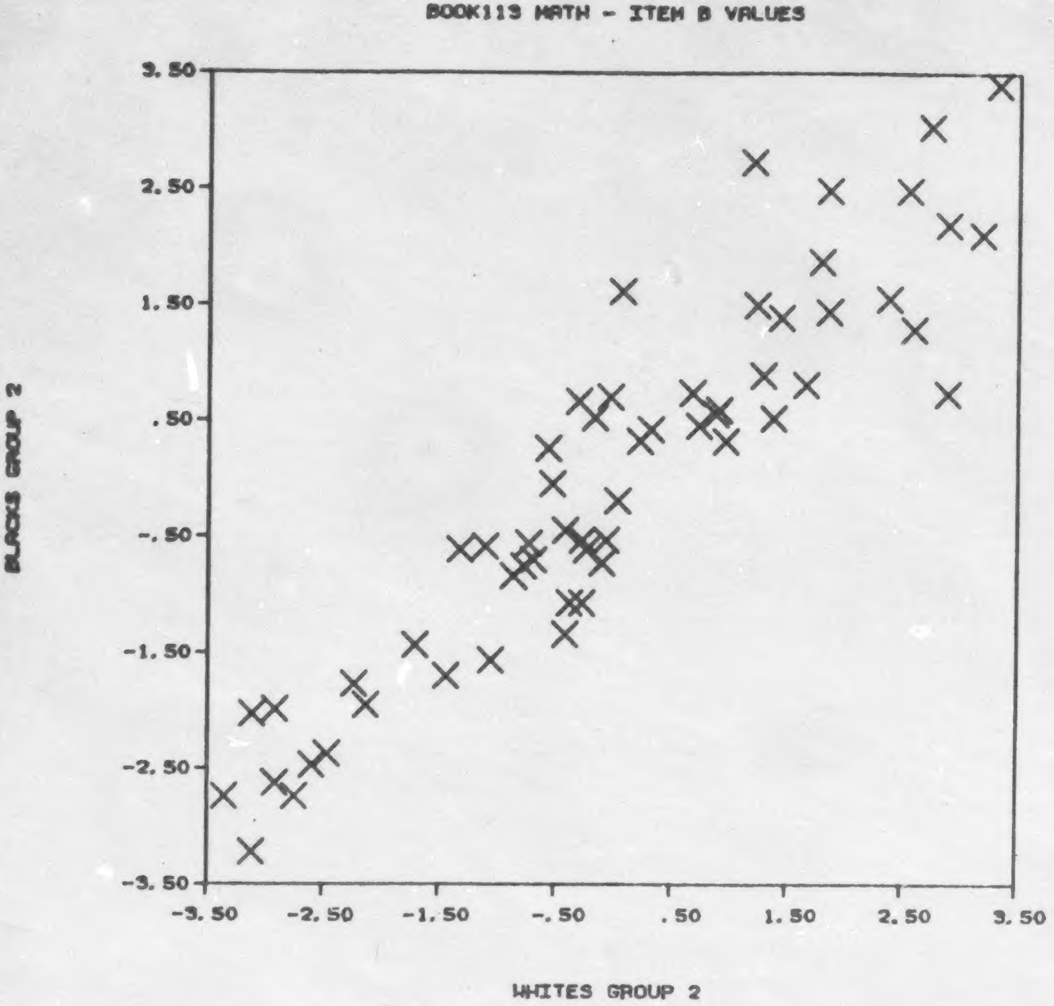


Figure 3.5.4 Plot of b values for the one-parameter model obtained from the second white and black student samples (N=165).

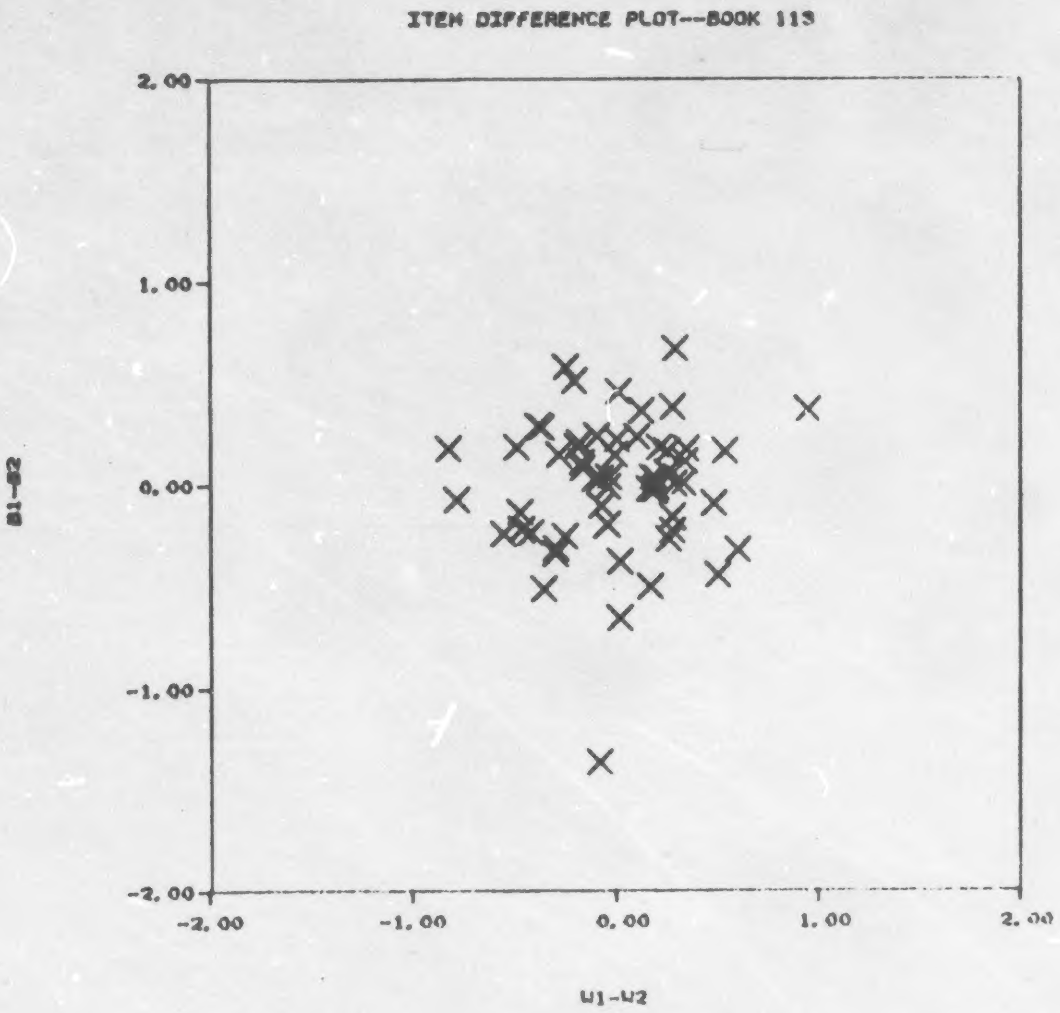


Figure 3.5.5 Plot of b value differences (Black 1 - Black 2 versus White 1 - White 2).

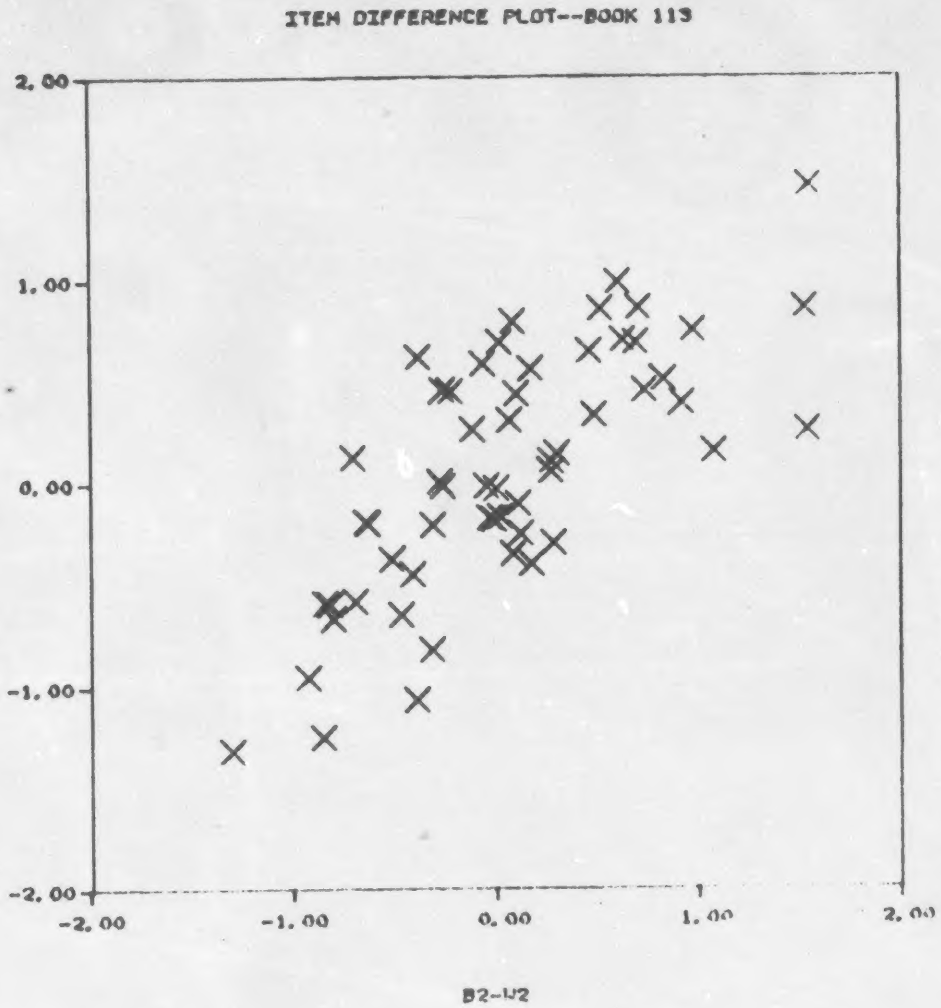


Figure 3.5.6 Plot of b value differences (Black 1 - White 1 versus Black 2 - White 2).

confounded with race (e.g., achievement scores/ability level - Blacks did perform substantially lower on the Math Booklets than Whites - see Table 3.4.2); and (2) failure to consider other important item statistics such as discrimination (a) and pseudo-chance level (c). With respect to (2), in other words, the problem is due to model-data misfit. But whatever the explanation it is clear that the feature of item parameter invariance is not obtained.[1]

In a follow-up investigation with NAEP Math Booklet No. 1 with 13 Year Olds, the examinee pool was split into high and low performers (the cut-off point was set at the median). Each group contained in excess of 1200 examinees. A one-parameter analysis was carried out with each group. Table 3.5.1 provides three difficulty estimates for each item: total group, low ability group, and high ability group. Plots of the three possible combinations are presented in Figures 3.5.7, 3.5.8, and 3.5.9.

The plots are not directly comparable with the earlier ones for race because the sample sizes in this analysis are considerably larger. But, again it seems clear that item parameter invariance is not obtained. This time, however, item parameter invariance was not obtained across high and low ability scorers. Table 3.5.2 and Figures 3.5.10, 3.5.11 and 3.5.12 provide a similar analysis for Math Booklet No. 2 with 13 Year Olds and again the conclusion is the same.

At least two criticisms can be made of the previous analyses summarized in Tables 3.5.1 and 3.5.2: (1) there is no baseline data available for interpreting the plots, and (2) no attempt is made to account for variation in items due to their discriminating power and pseudo-chance level. The analysis described next with Math Booklet No. 1 with 13 Year Olds was

---

[1] Unfortunately the same analyses could not be carried out with the three-parameter model because of the very small sample sizes. An alternate methodology to handle the small samples was recently proposed by Linn and Harnisch (1981).

Table 3.5.1

One-Parameter Model Difficulty Estimates for Total, Low, and High Ability Groups for NAEP Math Booklet No. 1 (13 Year Olds, 1977-78)

Item	Total	Ability Group	
		Low	High
1	-1.58	-1.47	-1.91
2	-2.45	-2.39	-2.64
3	-2.98	-2.88	-3.56
4	.45	1.00	-.08
5	-.19	.06	-.54
6	1.28	1.14	1.47
7	1.04	.89	1.26
8	-.48	-.84	.18
9	1.59	1.25	1.90
10	-1.22	-1.03	-1.87
11	-3.01	-2.88	-4.02
12	-2.64	-2.53	-3.24
13	-2.46	-2.32	-3.33
14	-2.38	-2.25	-3.07
15	-1.97	-1.87	-2.34
16	-1.89	-1.93	-1.61
17	-.65	-.83	-.20
18	2.78	3.10	2.82
19	-.05	.18	-.33
20	.08	.58	-.64
21	.27	.73	-.24
22	-2.60	-2.65	-2.23
23	1.25	.85	1.64
24	1.44	1.28	1.64
25	.47	.59	.45
26	-.67	-.83	-.25
27	3.13	3.10	3.27
28	2.13	1.34	2.68
29	2.37	2.51	2.44
30	.77	.93	.74
31	-.76	-.67	-.89
32	.18	.36	.03
33	.94	.79	1.16
34	-3.04	-3.10	-2.59
35	-.28	-.42	.04
36	2.14	1.01	3.01
37	1.24	1.07	1.46
38	-1.00	-.79	-1.71
39	-.48	-.80	.14
40	-.24	-.08	-.55

Table 3.5.1 (continued)

Item	Total	Abilty Group Low	High
41	1.79	2.05	1.79
42	-.43	-.39	-.42
43	-.99	-.92	-1.08
44	-.37	-.39	-.21
45	.80	.85	.87
46	1.37	1.66	1.32
47	3.02	2.01	3.64
48	2.67	2.65	2.80
49	.59	.79	.50
50	-1.47	-1.42	-1.50
51	2.40	1.73	2.83
52	.04	-.40	.69
53	-1.31	-1.42	-.91
54	-.67	-.63	-.63
55	1.91	2.52	1.83
56	-.62	-.59	-.56
57	1.54	2.33	1.33
58	-.77	-.71	-.80

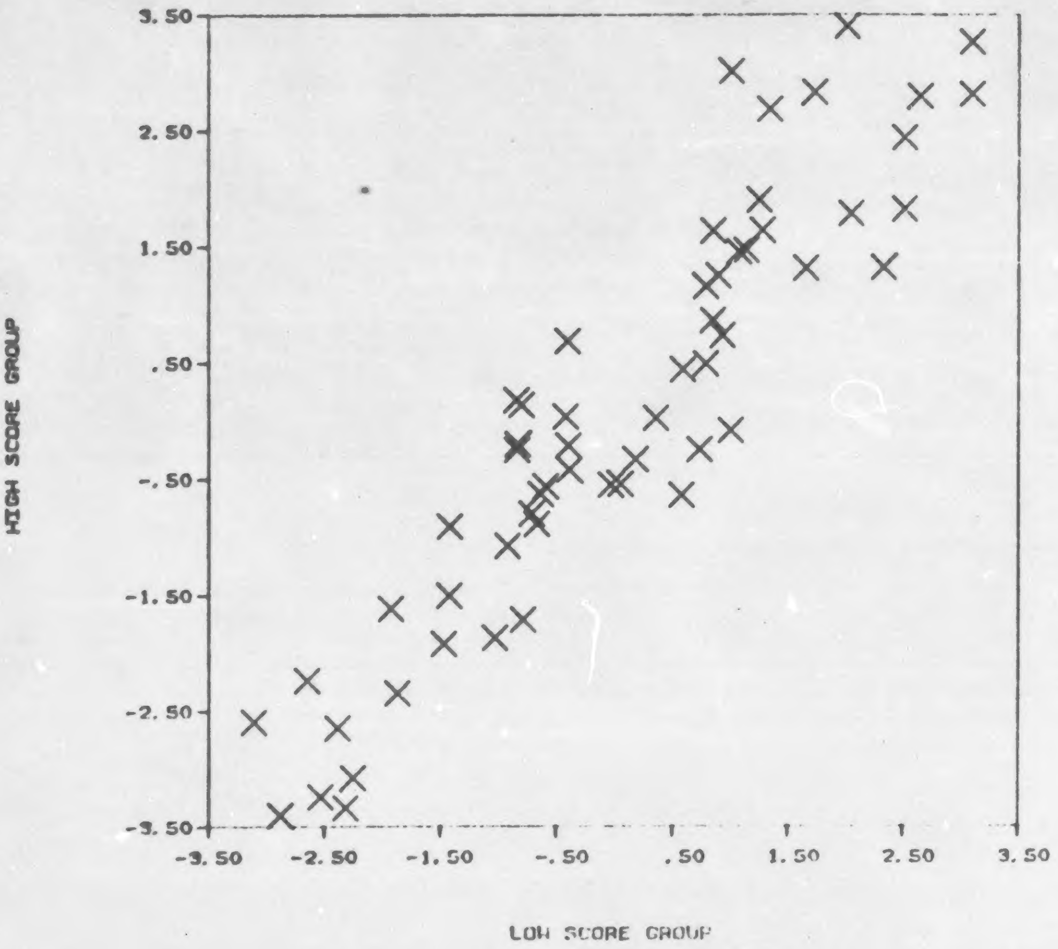


Figure 3.5.7 Plot of one-parameter model item difficulty estimates for the low and high scoring ability groups on NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

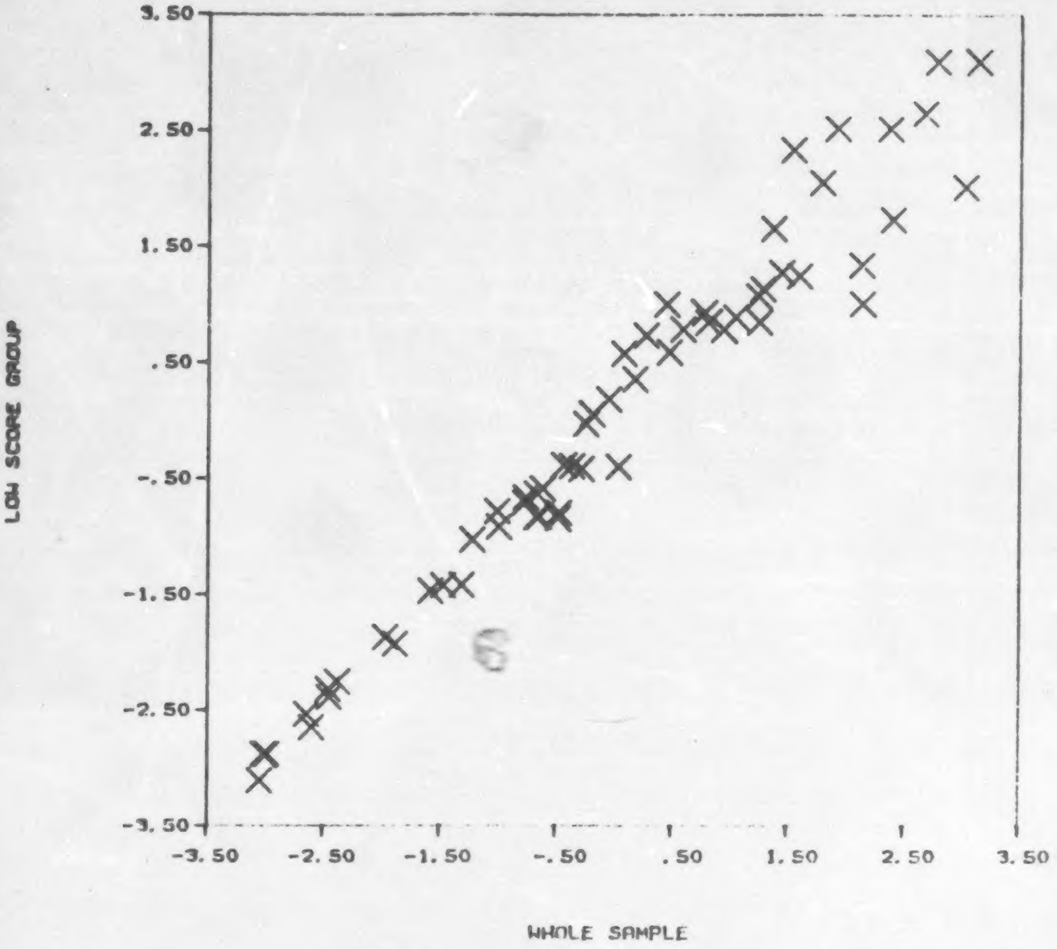


Figure 3.5.8 Plot of one-parameter model item difficulty estimates for the low and total ability groups on NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

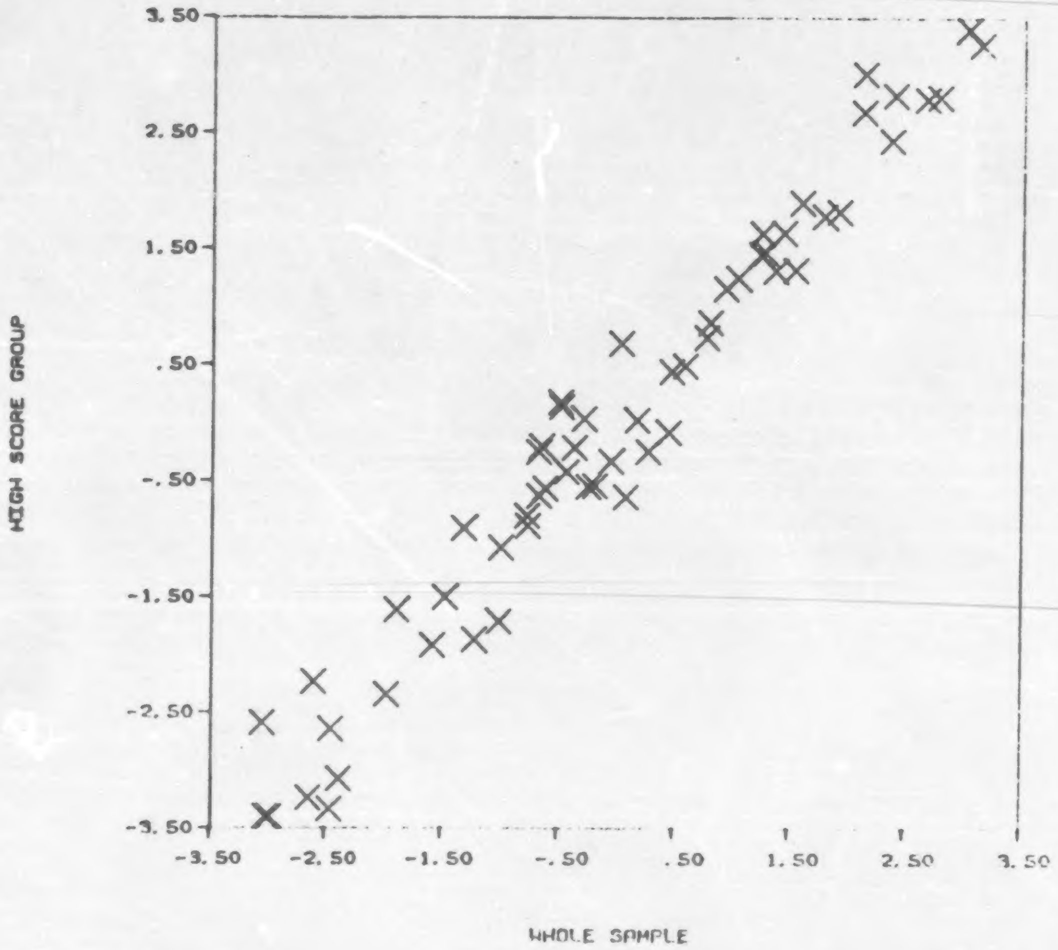


Figure 3.5.9 Plot of one-parameter model item difficulty estimates for the high and total ability groups on NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

Table 3.5.2

One-Parameter Model Difficulty Estimates for Total, Low, and High Ability Groups for NAEP Math Booklet No. 2 (13 Year Olds, 1977-79)

Item	Total	Ability Group Low	High
1	.21	.21	.28
2	.72	.91	.63
3	-.22	-.42	.14
4	-.45	-.63	-.09
5	.24	.43	.06
6	2.52	2.80	2.58
7	.57	.60	.61
8	-3.23	-3.26	-2.99
9	-1.58	-1.36	-2.91
10	-1.51	-1.29	-2.73
11	-1.37	-1.16	-2.36
12	-1.12	-.89	-1.97
13	-2.49	-2.41	-2.85
14	1.00	.75	1.26
15	1.72	1.98	1.68
16	-2.05	-2.02	-2.06
17	-1.58	-1.52	-1.76
18	-1.72	-1.64	-1.95
19	-3.13	-3.25	-2.48
20	-3.12	-3.12	-2.98
21	-2.96	-3.06	-2.36
22	3.92	2.35	4.79
23	1.26	1.70	1.09
24	.87	1.19	.70
25	.02	.05	.04
26	-1.87	-2.23	-.93
27	.37	-.10	.92
28	2.06	1.96	2.17
29	-2.34	-2.34	-2.22
30	2.66	2.78	2.70
31	1.67	2.05	1.59
32	-.32	-.55	-.11
33	-3.01	-2.88	-4.01
34	-1.72	-1.70	-1.73
35	2.19	2.42	2.19
36	.13	.17	.15
37	-.32	-.06	-.78
38	1.44	1.56	1.45
39	.61	.67	.62
40	.86	.98	.83

Table 3.5.2 (continued)

Item	Total	Ability Group Low	High
41	-1.86	-1.79	-2.12
42	-1.03	-.92	-1.32
43	-1.01	-1.01	-.93
44	-.67	-.97	-.08
45	-.26	-.36	-.04
46	-1.25	-1.19	-1.35
47	-1.18	-1.01	-1.73
48	1.91	2.14	1.90
49	2.61	3.32	2.55
50	-.09	.13	-.39
51	1.57	.54	2.32
52	2.60	1.90	2.93
53	1.19	1.76	.96
54	-1.69	-1.80	-1.28
55	.75	.66	.88
56	.62	1.11	.26
57	3.00	1.78	3.64
58	2.62	2.82	2.64
59	2.46	2.13	2.65
60	-.82	-.89	-.59
61	-.15	-.28	.80
62	1.77	2.26	1.66

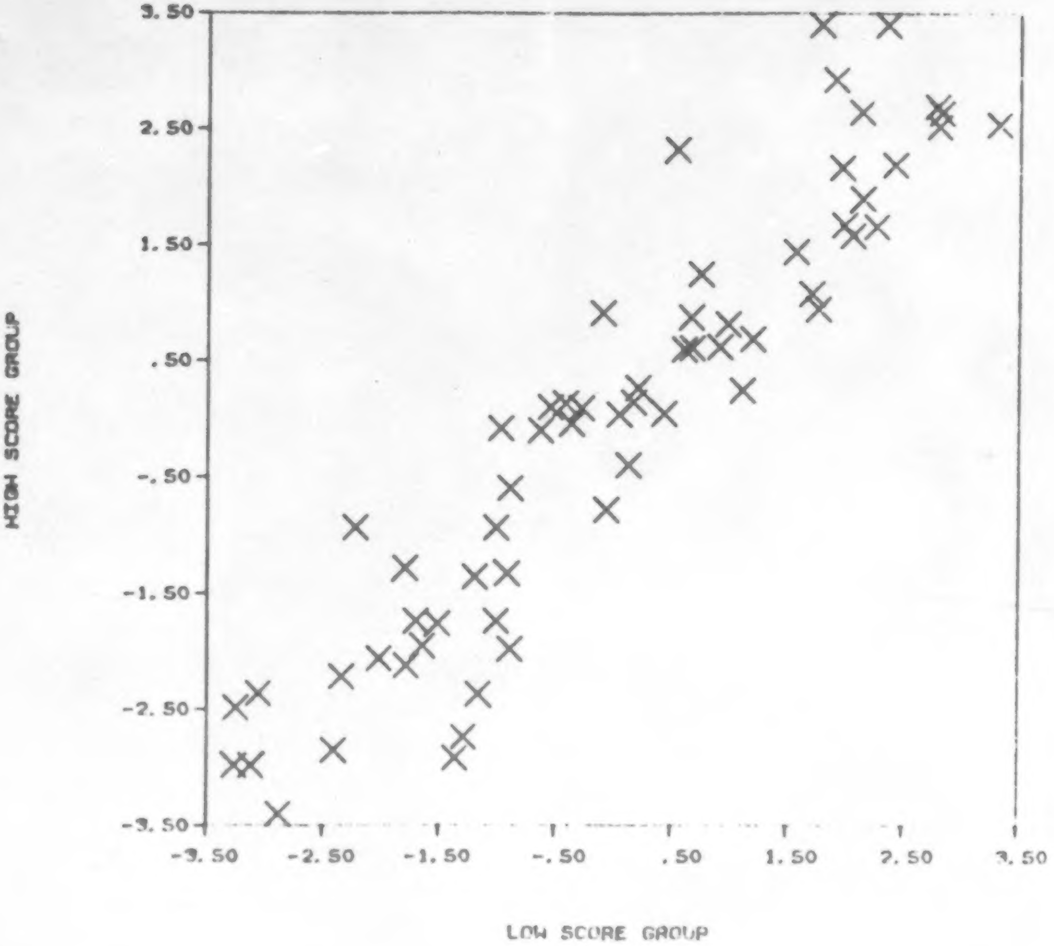


Figure 3.5.10 Plot of one-parameter model item difficulty estimates for the low and high scoring ability groups on NAEP Math Booklet No. 2 (13 Year Olds, 1977-78).

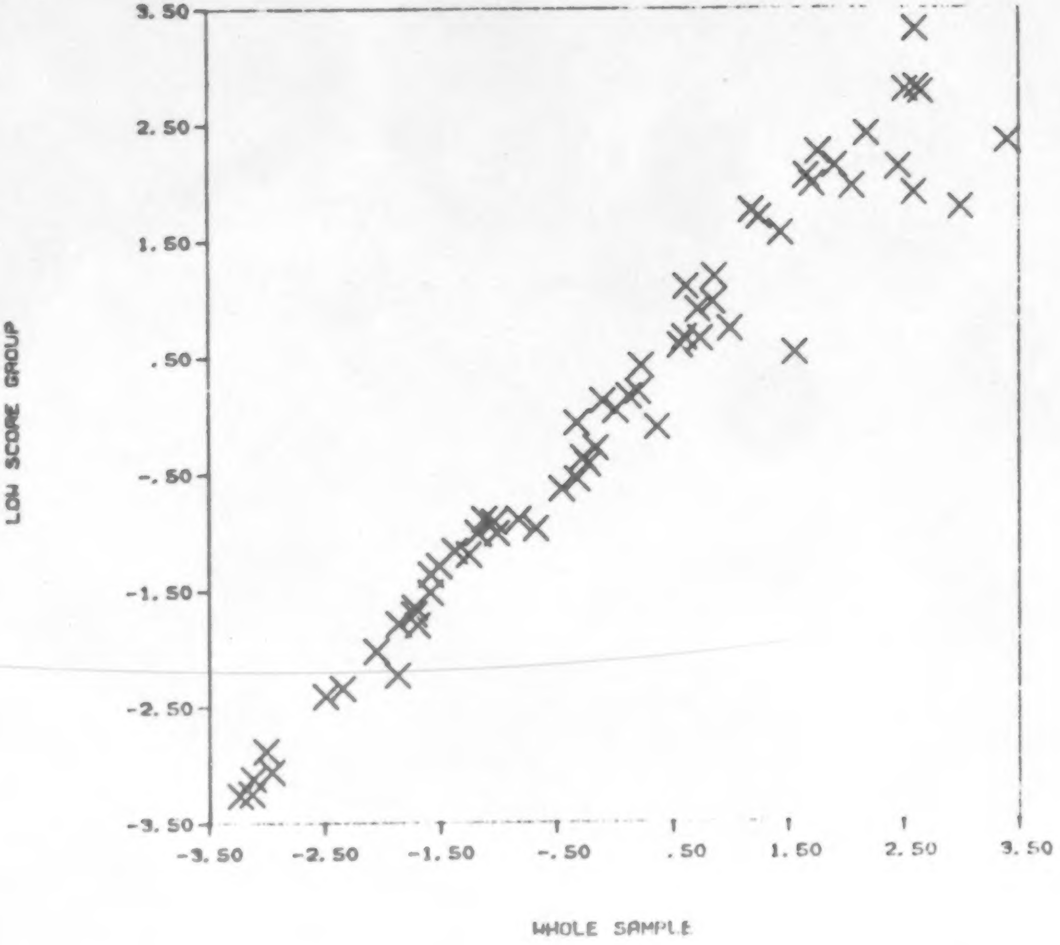


Figure 3.5.11 Plot of one-parameter model item difficulty estimates for the low and total ability groups on NAEP Math Booklet No. 2 (13 Year olds, 1977-78).

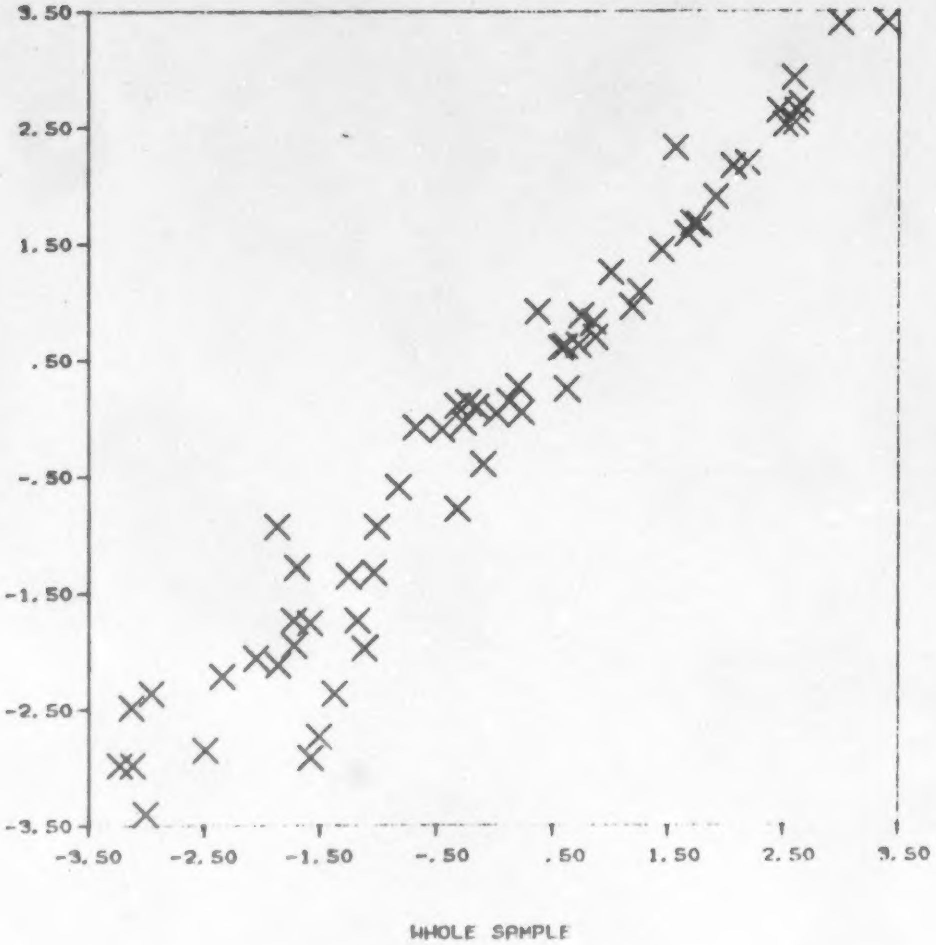


Figure 3.5.12 Plot of one-parameter model item difficulty estimates for the high and total ability groups on NAEP Math Booklet No. 2 (13 Year Olds, 1977-78).

carried out to address the two deficiencies. A group of 2400 examinees was found with the 1200 lowest ability students and 1200 highest ability students. The (approximately) 22 middle ability students were deleted from the analysis. Next, the 2400 examinees were divided on a random basis into two equal sub-groups of 1200 examinees. Each sub-group was used to obtain one-parameter and three-parameter model item estimates. Figures 3.5.13 and 3.5.14 provide the plots of b values in the two samples obtained with the one- and three-parameter logistic models. The item parameter estimates in the two samples with either test model are nearly identical. Thus, item parameter invariance across random groups is established. Next, the 2400 examinees were divided into two equal-sized low and high ability groups (again, N=1200) and the analyses and plots carried out with the random groups were repeated. The results for the one- and three-parameter models are reported in Figures 3.5.15 and 3.5.16 respectively.

If the feature of item invariance was present all four plots should have looked the same. In fact, the plots in Figures 3.5.15 and 3.5.16 are substantially different from those in Figure 3.5.13 and 3.5.14. However, it is not plausible at this time to explain the differences in terms of a failure to account for essential item statistics (i.e., discrimination and pseudo-level) since the one-parameter and three-parameter plots of item difficulties for high and low ability examinees shown in Figures 3.5.15 and 3.5.16 are similar. One possible explanation which remains is that item parameter estimation is not done very well when extreme groups are used.[1] Of course another possibility is that the test items are functioning differently in the two ability groups, i.e., item parameters are not invariant across ability groups.

---

[1] The close fit between the three-parameter model and several data sets reported in section 3.6 suggest that this explanation is highly plausible.

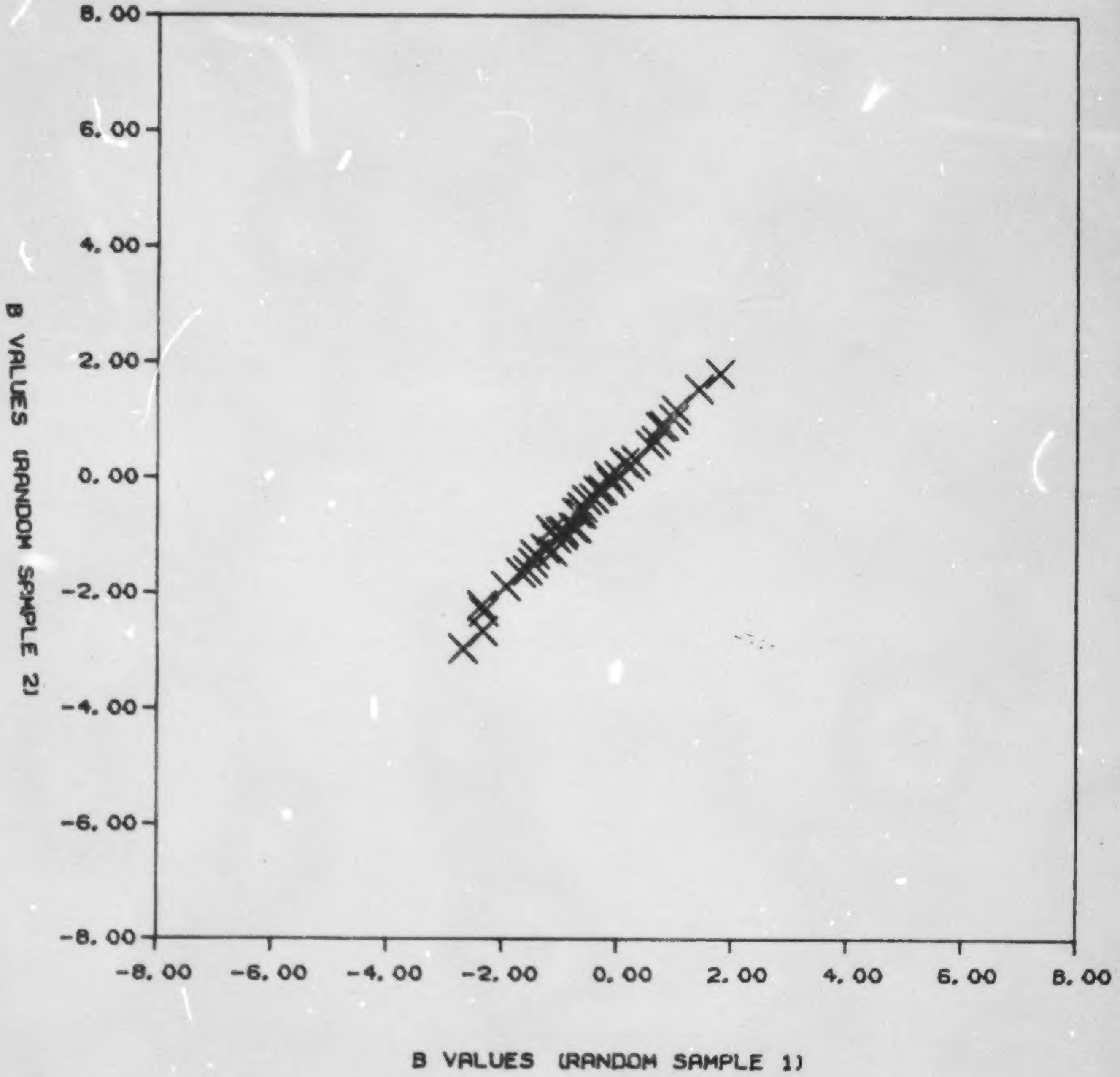


Figure 3.5.13. Plot of one-parameter model item difficulty estimates obtained in two equivalent samples with NAEP Math Booklet No. 1 (13 Year Olds, 1977-78, N=1200).

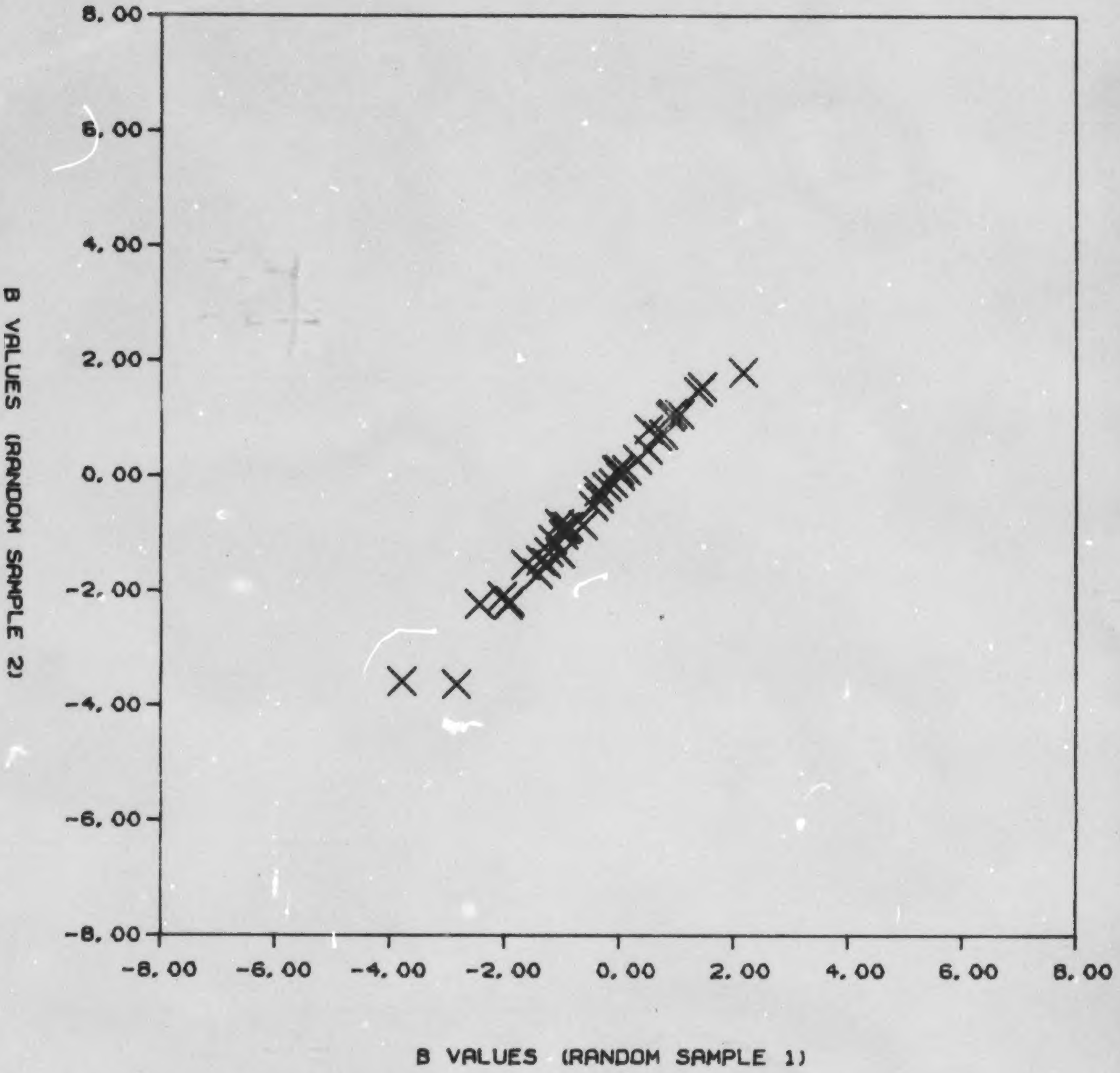


Figure 3.5.14. Plot of three-parameter model item difficulty estimates obtained in two equivalent samples with NAEP Math Booklet No. 1 (13 Year Olds, 1977-78, N=1200).

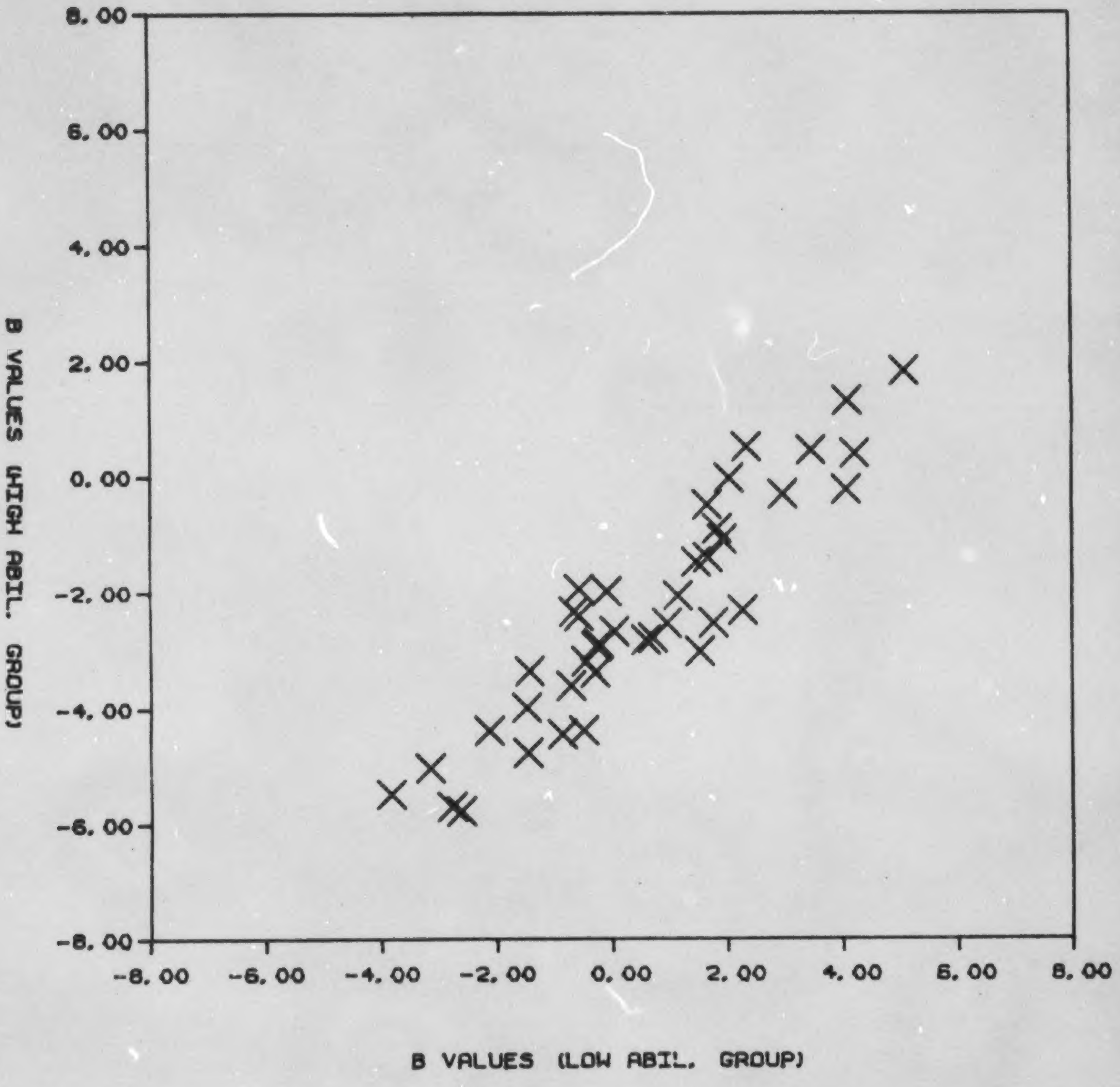


Figure 3.5.15. Plot of one-parameter model item difficulty estimates obtained in low and high ability groups with NAEP Math Booklet No. 1 (13 Year Olds, 1977-78, N=1200).

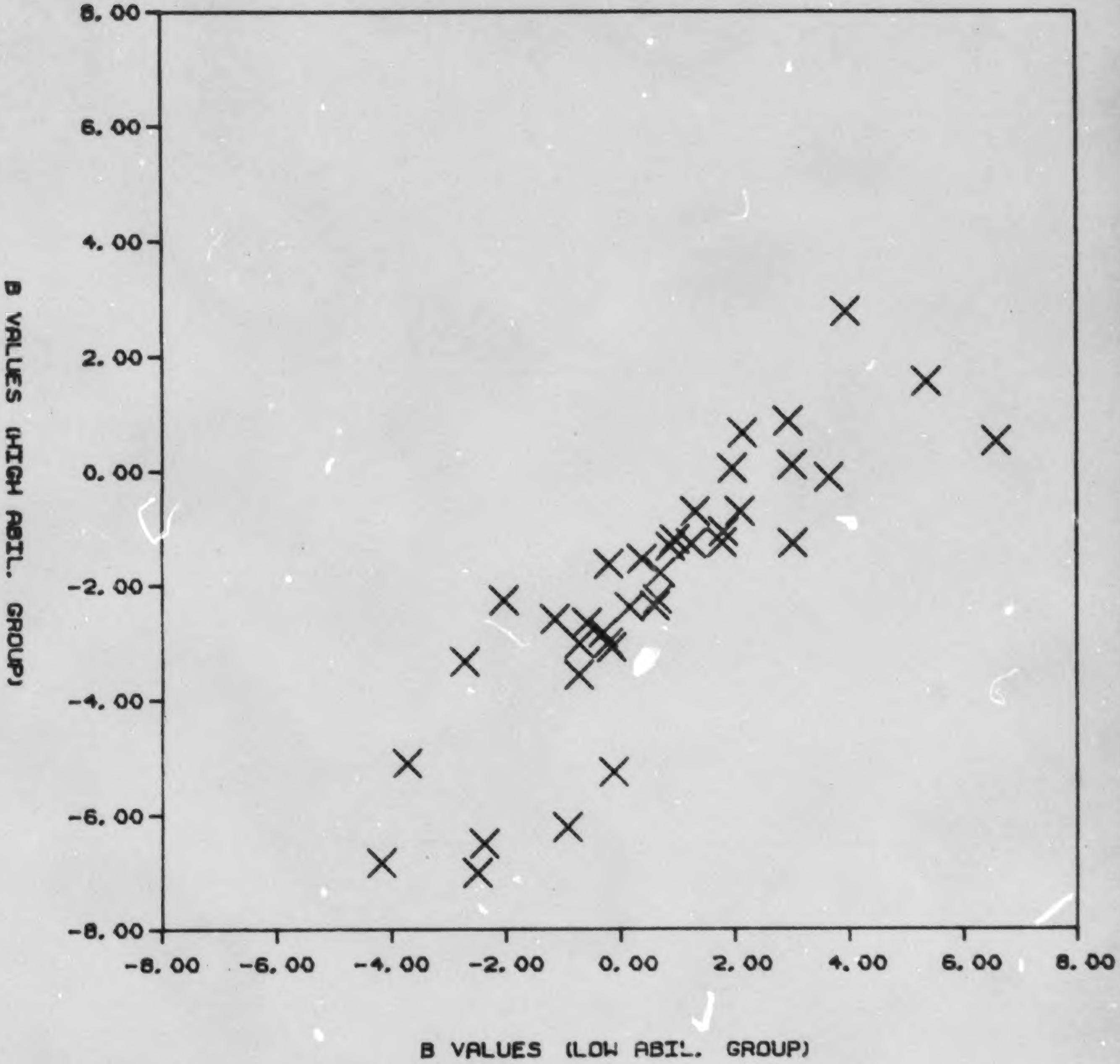


Figure 3.5.16. Plot of three-parameter model item difficulty estimates obtained in low and high ability groups with NAEP Math Booklet No. 1 (13 Year Olds, 1977-78, N=1200).

Tables 3.5.3 and 3.5.4 provide the results from a different type of analysis but one which seems promising for addressing item parameter invariance. It has become common practice to address item invariance by conducting a statistical analysis of the b value differences between two groups of examinees (for example, Blacks and Whites). The problem is, as was stated in section 2.1, when the sample size is large, even practically insignificant differences are often statistically significant. The method described next depends upon replication and practically significant differences. In this analysis, a practically significant difference of interest (referred to as the critical value) was selected (=.50) and two equal-sized Black and White examinee samples were divided into two equal-sized sub-samples (N=165). A one-parameter analysis of Math Booklet No. 1 with 13 Year Olds was carried out for each group. In Table 3.5.3 items with b value differences exceeding .50 between either the first Black and White samples, or the second Black and White samples, are shown. Twenty-seven of 58 items exceeded the critical value with the first samples; 25 of the 58 items exceeded the critical value with the second samples. If the differences were due to chance factors only, 20.0% of the test items would be predicted to be identified in both samples. In fact, 59.2% of the 27 items (16 items) identified as different in the first samples were identified as different in the second samples. When viewed in the other direction, 16 of the 25 items (or 80%) identified in the second samples were also identified in the first samples.

Table 3.5.4 provides the results of a replication of the study with Math Booklet No. 2 for 13 Year Olds. Eighteen items were identified in sample 1; 15 items were identified in sample 2; 13 of the items were common. In other words, of the 18 items identified in the first samples, 72.2% were

Table 3.5.3

Item Difficulty Differences in Black-White Samples for NAEP Math Booklet No. 1[1] (13 year olds, 1977-78)

Item	Black and White Samples (N=165) 1	2
4	1.40	1.55
6	-1.06	
7	- .82	
8	- .95	- .92
10	.86	.52
11	.99	.61
12	.80	
13		1.08
15	.65	
19	.69	.70
20	.76	.98
21		.74
23	-1.25	- .85
24	- .59	- .84
26	- .58	- .69
27		1.54
28	-1.31	-1.30
29	.59	
33		- .64
35		- .63
36	-1.83	-2.15
38	.86	.71
39	- .66	- .80
40	.51	.83
43	.57	
47	-2.03	-1.08
48		- .70
51	- .59	- .82
52	- .65	
53		- .51
55	.71	.63
56	.62	
57	.86	1.53
58	.70	

[1] Only items with a difference $\geq |.50|$ in one or both samples are reported.

Table 3.5.4

Item Difficulty Differences in Black-White Samples for NAEP Math Booklet No. 2[1] (13 year olds, 1977-78)

Item	Black and White Samples (N=220) 1	2
4		- .57
6	.51	
9	1.11	.88
10	1.45	.66
11	.98	.88
12	.86	
13	.70	
15	.53	
19		.62
20	1.27	
22	-2.03	-1.95
26	- .75	-1.36
29	.58	.49
30	.54	.83
35	1.03	.76
50	.62	.62
51	-1.33	-1.18
53	.63	.94
57	-1.84	-1.54
61	- .84	- .65

[1] Only items with a difference ≥ |.50| in one or both samples are reported.

identified in the second samples; and of the 15 items identified in the second samples, 86.7% of the items were identified in the first samples. If chance factors only were operating, about 7% of the items would be expected to be commonly identified in the two samples (15/62 x 18/62). Clearly, it cannot be argued that item invariance across the two groups is present when the b values are estimated using the one-parameter logistic test model.

The method described above seems like a promising approach for addressing the problem of item invariance. And, from one point of view, it really doesn't matter what the causes of the differences are. The fact is that item invariance is not found across a variable that can be used to describe the examinee population. It would be misleading therefore to offer only a single set of item statistics. While not investigated here, commonly identified items can be additionally studied to attempt to detect the source of the problem(s). At this stage, directions of any observed differences can be also investigated.

3.6 Checking Additional Model Predictions

This section of our work is divided into two parts: Residual Analyses, and Research Hypothesis Investigations.

Residual Analyses

To carry out residual analyses with Math Booklets Nos. 1, 2, and 3 for 9 and 13 Year Olds it was necessary to prepare a computer program. A listing and sample output of our program is presented in Appendix B. The program was prepared to be compatible with the item and ability parameter estimation output from LOGIST. The program provides both residuals and standardized residuals for each test item at various ability levels (the number is selected by the user). (Twelve ability levels were chosen in our investigation.) In addition, fit statistics are available for each test item (found by summing over ability levels), for each ability level (found by summing over test items), and for the total test (found by summing over ability levels and test items).

A sample set of standardized residuals for Math Booklet No. 1 with 13 Year Olds[1] obtained with the one-parameter model are shown in Figures 3.6.1 to 3.6.11. Two features of the plots in the figures (and other plots we studied) are the cyclic patterns and the large size of the standardized residuals. Item patterns like those in Figures 3.6.1, 3.6.3, 3.6.4, 3.6.5, and 3.6.10 were obtained for items with relatively high biserial correlations. Item patterns like those in Figures 3.6.6, 3.6.7, 3.6.8, and 3.6.9 were obtained for items with relatively low biserial correlations. Also, the standardized residuals tended to be high. In Table 3.6.1 it can be seen that (approximately) 25% of the standardized residuals exceeded a value of 3 when the one-parameter model was fit to the test data. This

---

[1] Standardized residual plots for items 1 to 10, and 36 are shown in the figures.

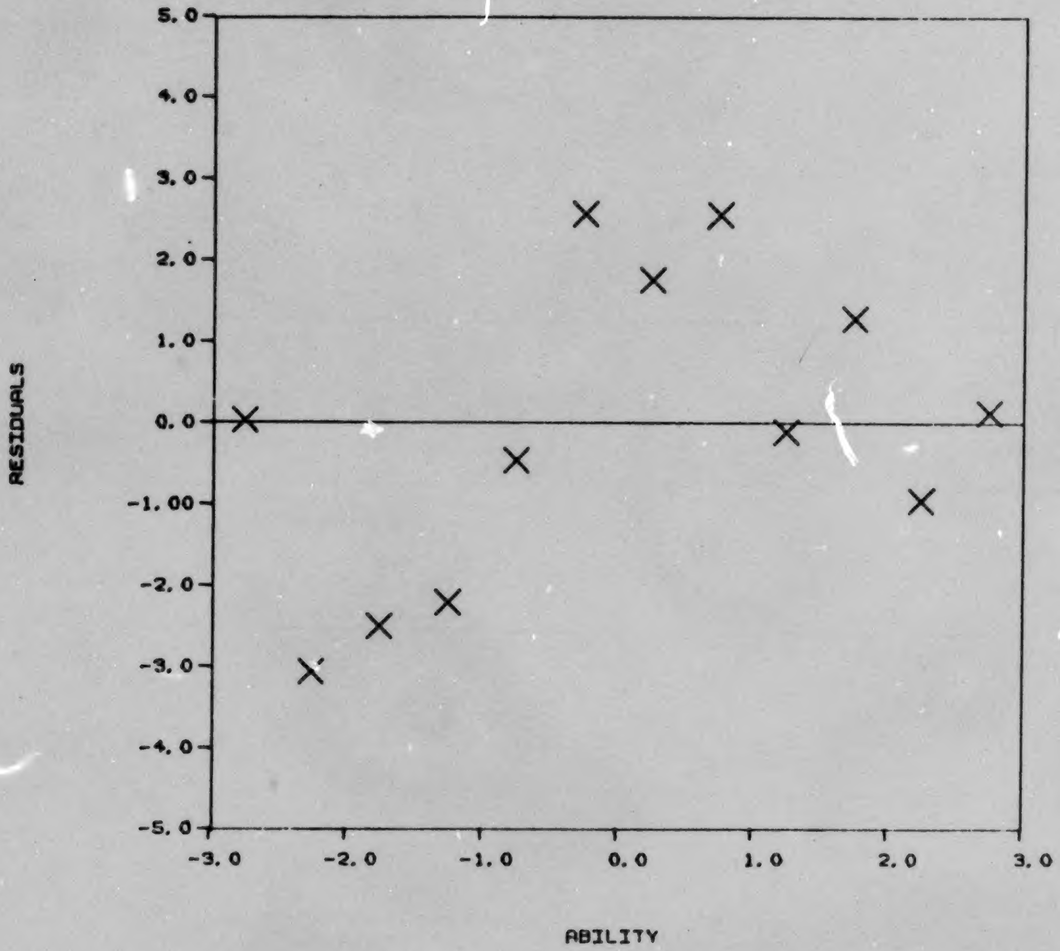


Figure 3.6.1 Standardized residual plot obtained with the one-parameter model for test item 1 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

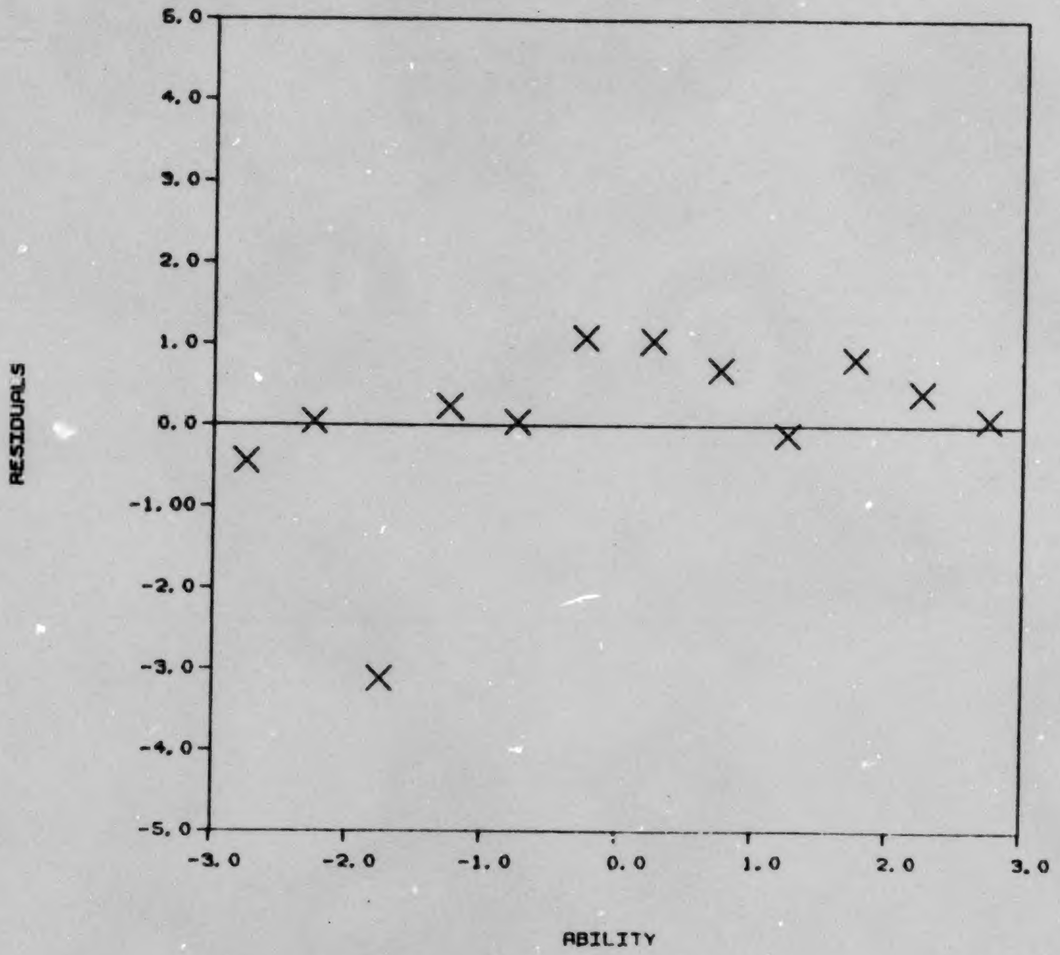


Figure 3.6.2 Standardized residual plot obtained with the one-parameter model for test item 2 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

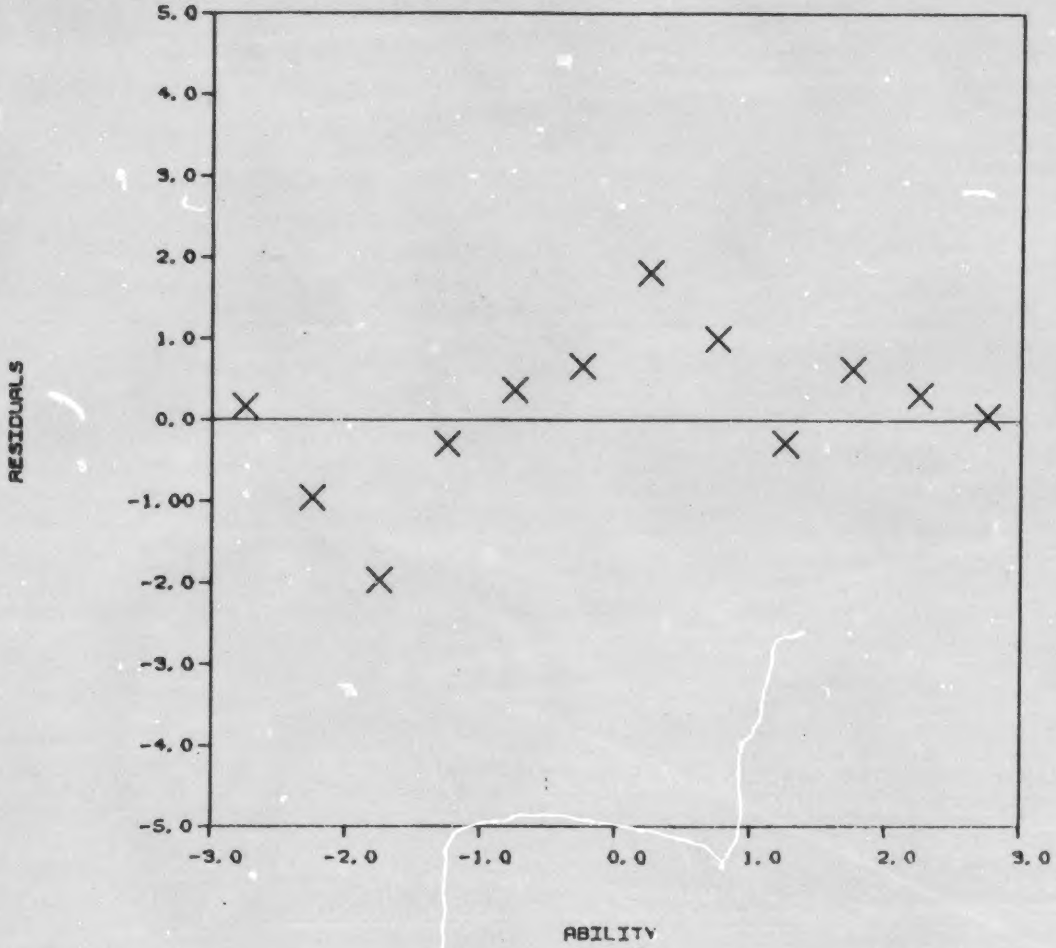


Figure 3.6.3 Standardized residual plot obtained with the one-parameter model for test item 3 from NAEP Math Booklet No. 1 (13 Year olds, 1977-78).

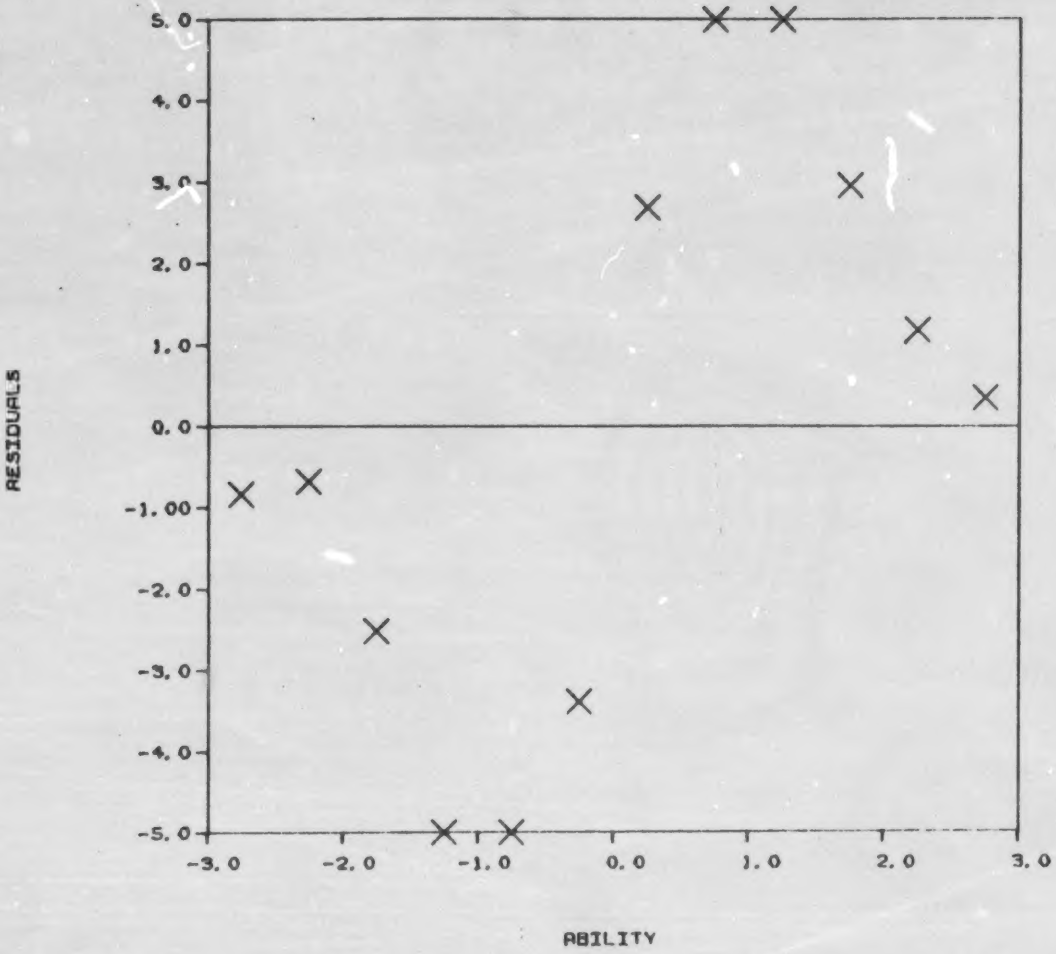


Figure 3.6.4 Standardized residual plot obtained with the one-parameter model for test item 4 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

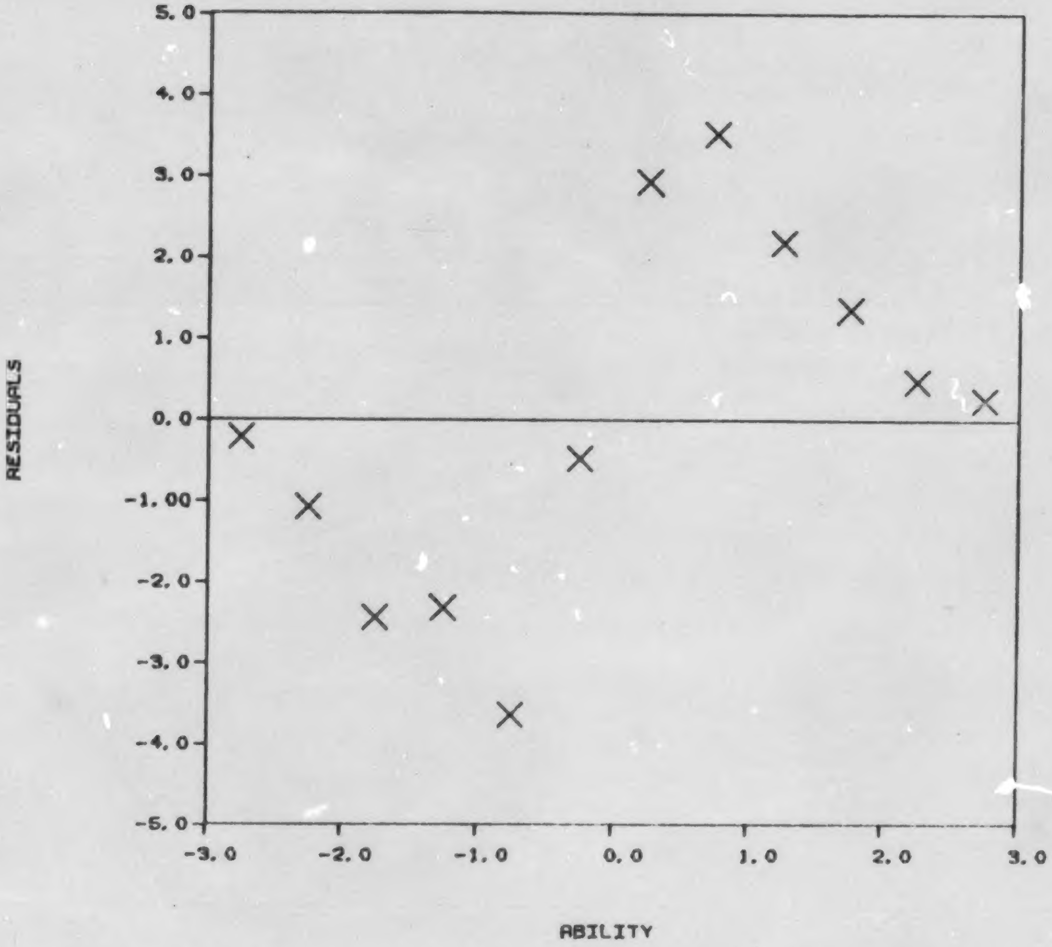


Figure 3.6.5 Standardized residual plot obtained with the one-parameter model for test item 5 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

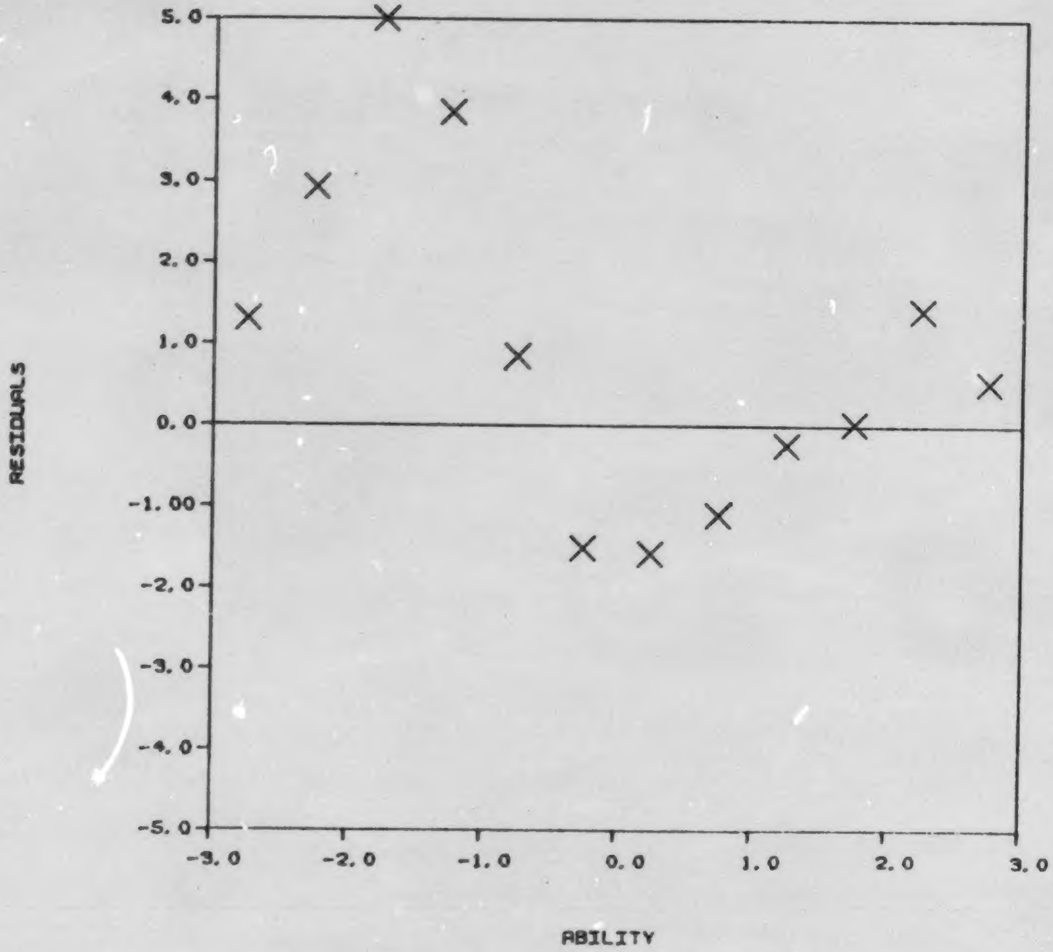


Figure 3.6.6 Standardized residual plot obtained with the one-parameter model for test item 6 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

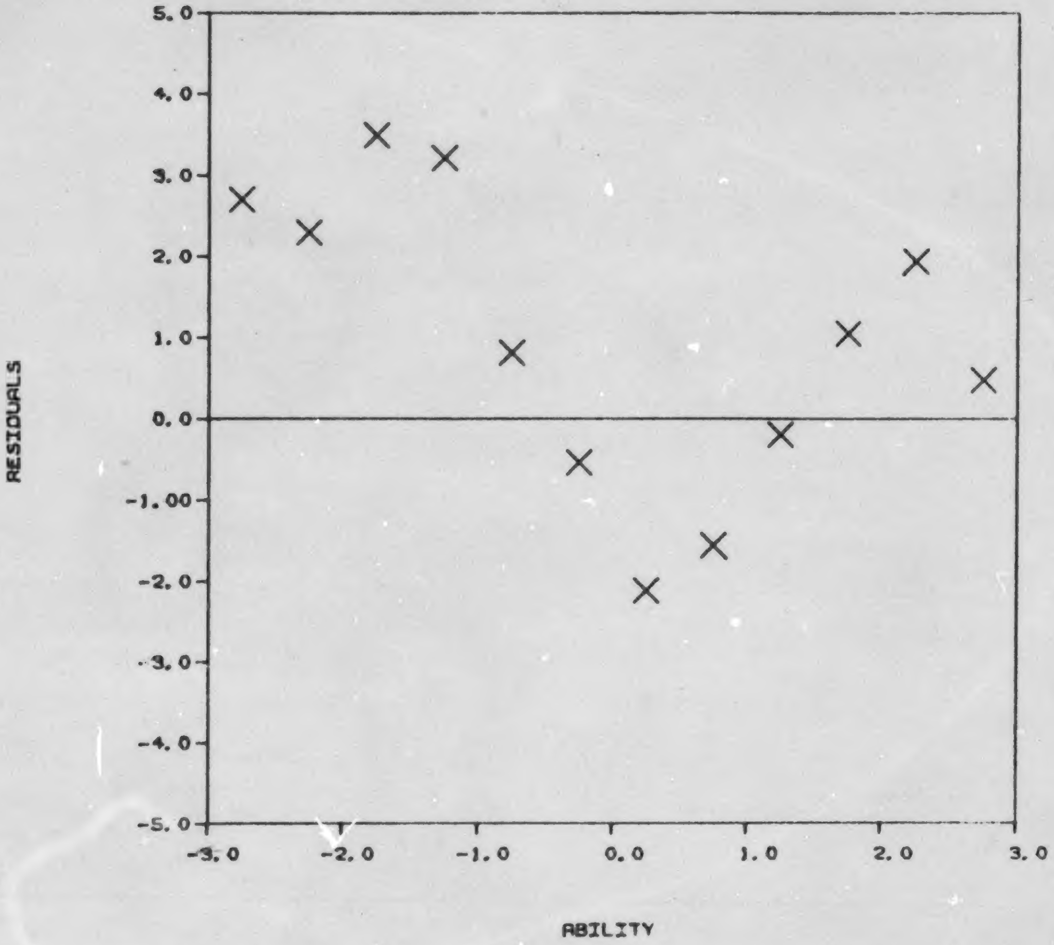


Figure 3.6.7 Standardized residual plot obtained with the one-parameter model for test item 7 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

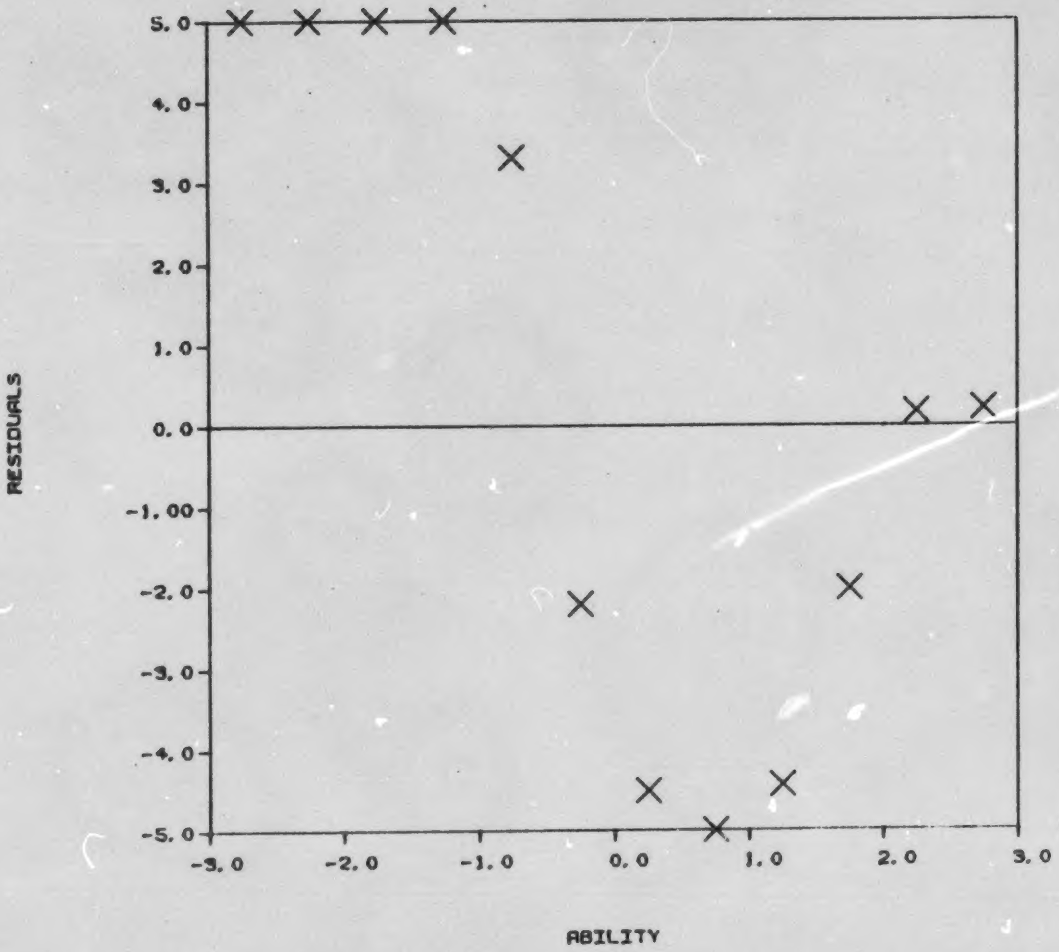


Figure 3.6.8 Standardized residual plot obtained with the one-parameter model for test item 8 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

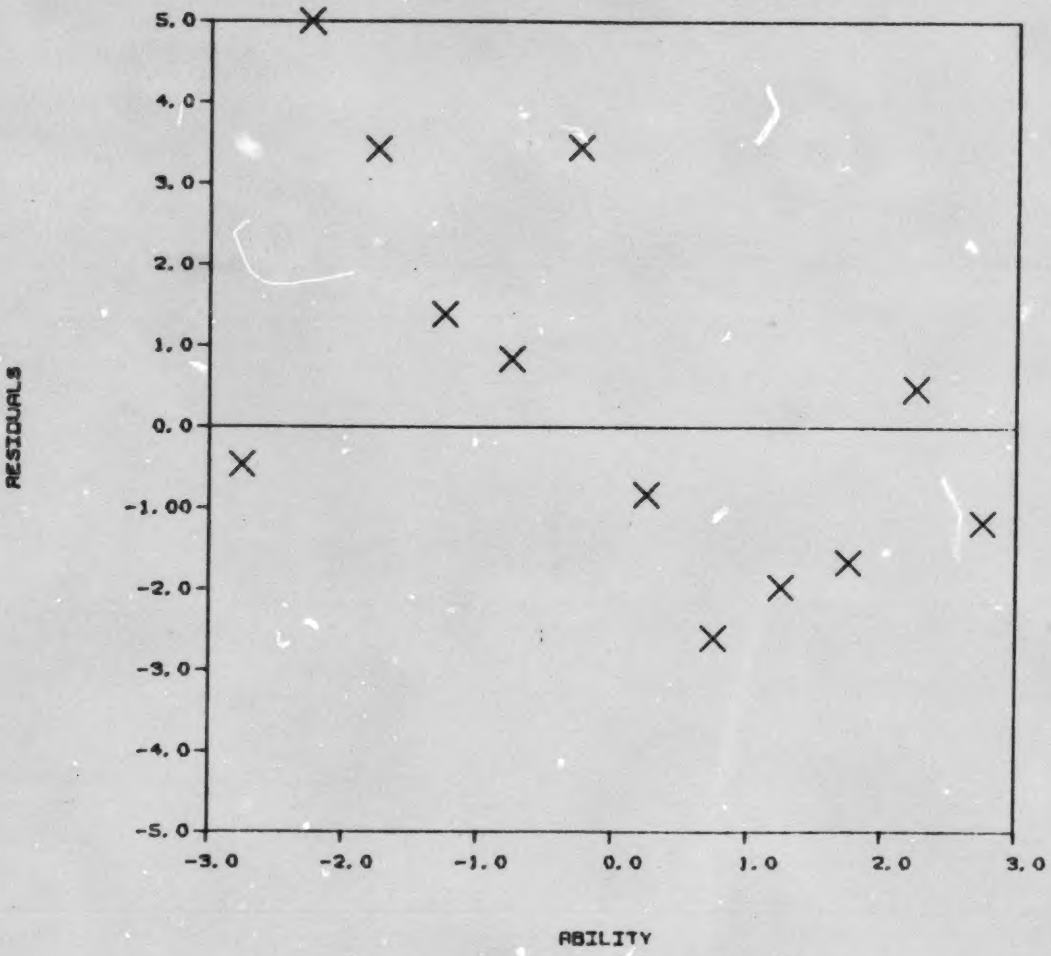


Figure 3.6.9 Standardized residual plot obtained with the one-parameter model for test item 9 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

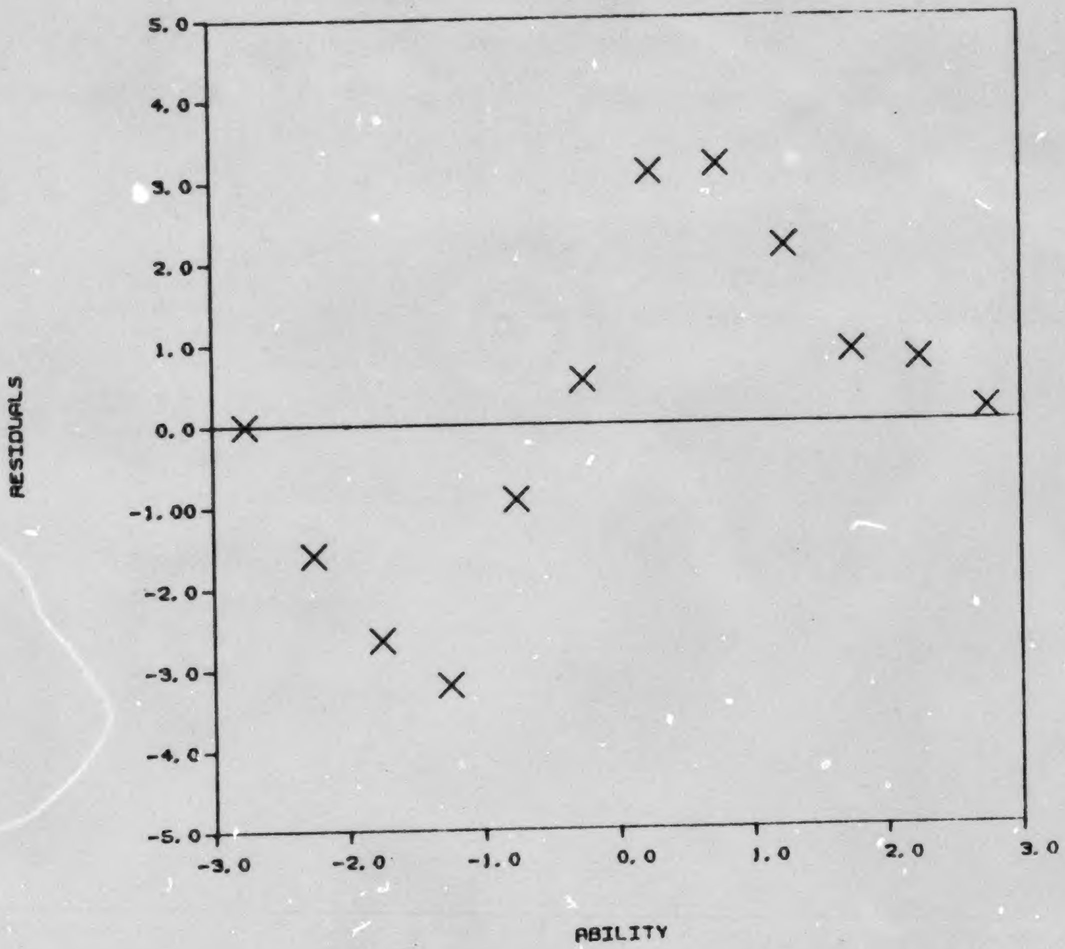


Figure 3.6.10 Standardized residual plot obtained with the one-parameter model for test item 10 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

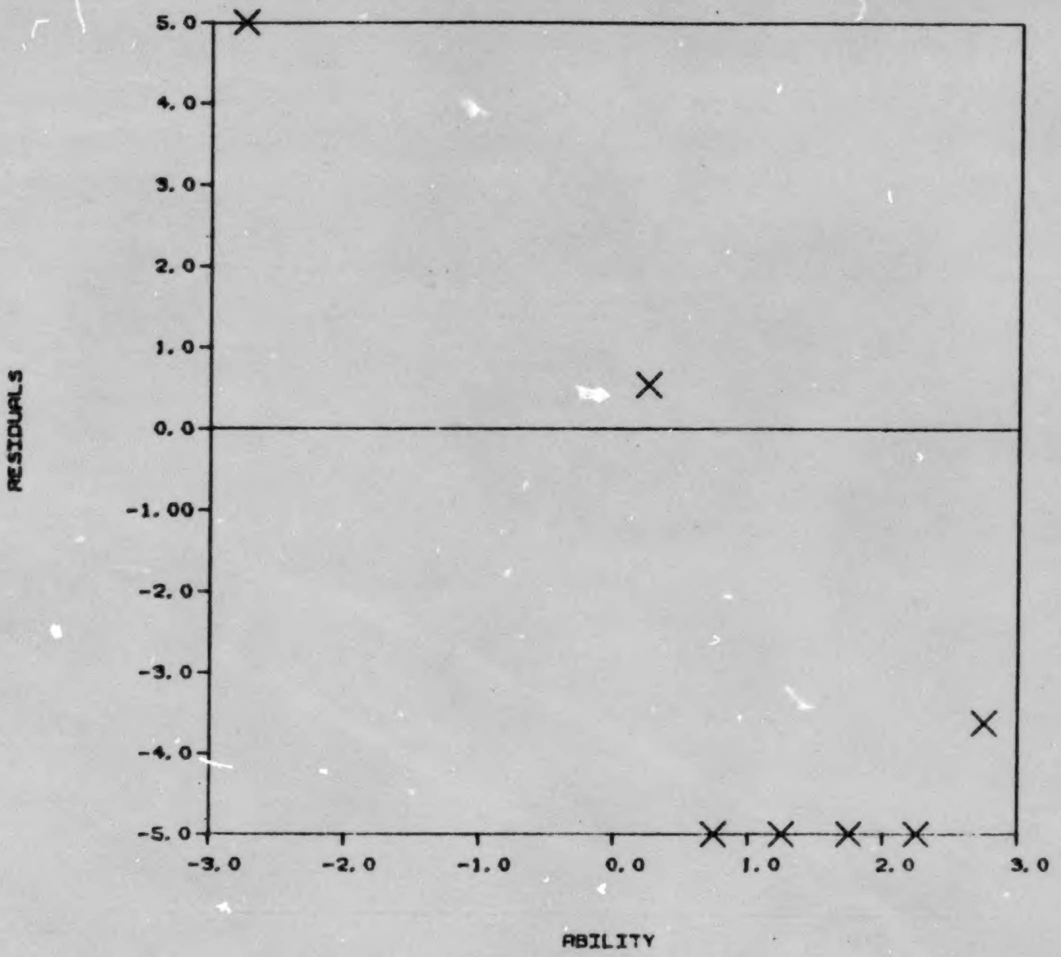


Figure 3.6.11 Standardized residual plot obtained with the one-parameter model for test item 36 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

result was obtained with 6 test booklets. If the model data fit had been good, the distribution of standardized residuals would have been approximately normal.

The standardized residual plots obtained from fitting the three-parameter model and shown in Figures 3.6.12 to 3.6.22 reveal dramatically different patterns. The cyclic patterns which were so evident in the first eleven figures are gone, and the sizes of the standardized residuals are substantially smaller.

Table 3.6.1 provides a complete summary of the distributions of standardized residuals obtained with the one- and three-parameter models for six Math Booklets. In all cases the standardized residuals are considerably smaller with the three-parameter model and the distributions are approximately normal.

Table 3.6.2 reports the average raw and absolute-valued standardized residuals[1] at 12 ability levels with the one- and three-parameter models for the same six Math Booklets. Again, the results in this table reveal the superiority of the three-parameter model. Also, it is clear that the three-parameter model is especially effective at low levels of ability.

Research Hypothesis Investigations

The residual analysis results in the last section were most interesting but it seemed desirable to investigate the misfit statistics further. Tables 3.6.3 to 3.6.6 provide the basic information we worked with for four of the Math Booklets.

---

[1] The average raw standardized residuals provide information about the size and direction of the misfit between the observed results and the ICCs.

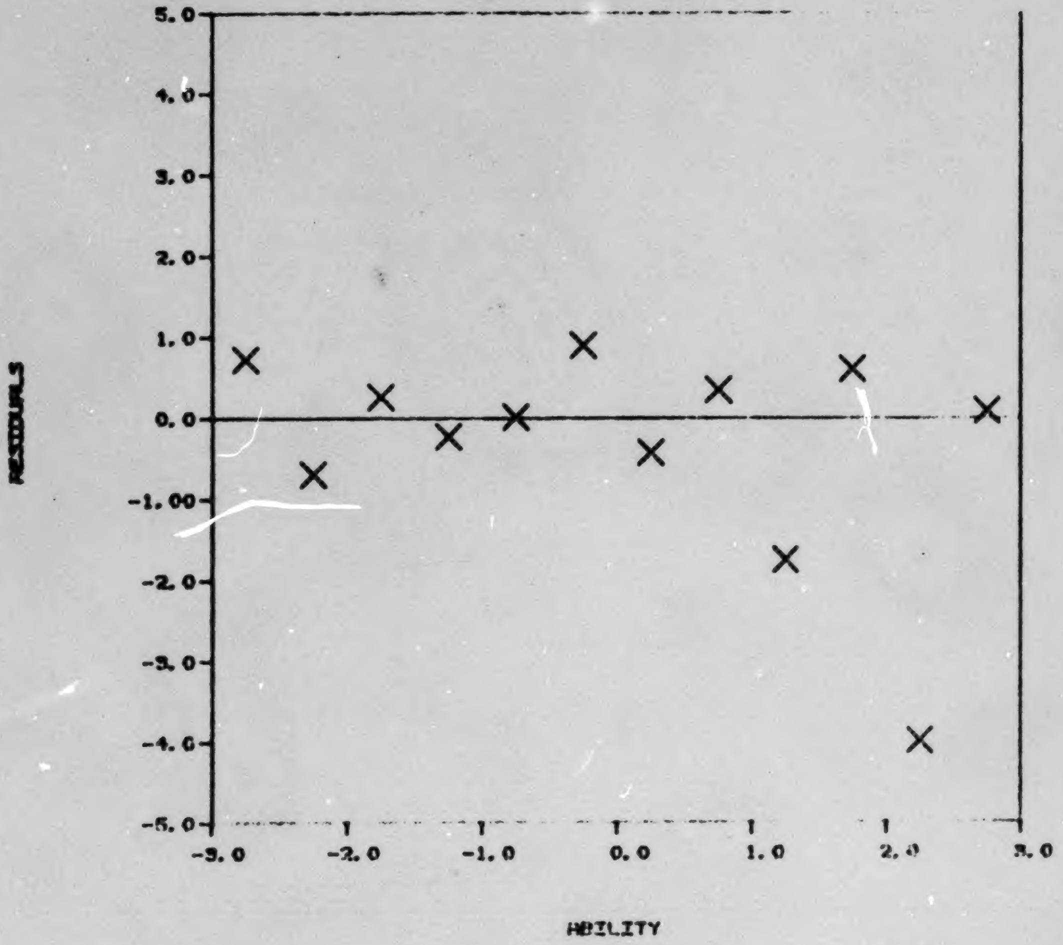


Figure 3.6.12 Standardized residual plot obtained with the three-parameter model for test item 1 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

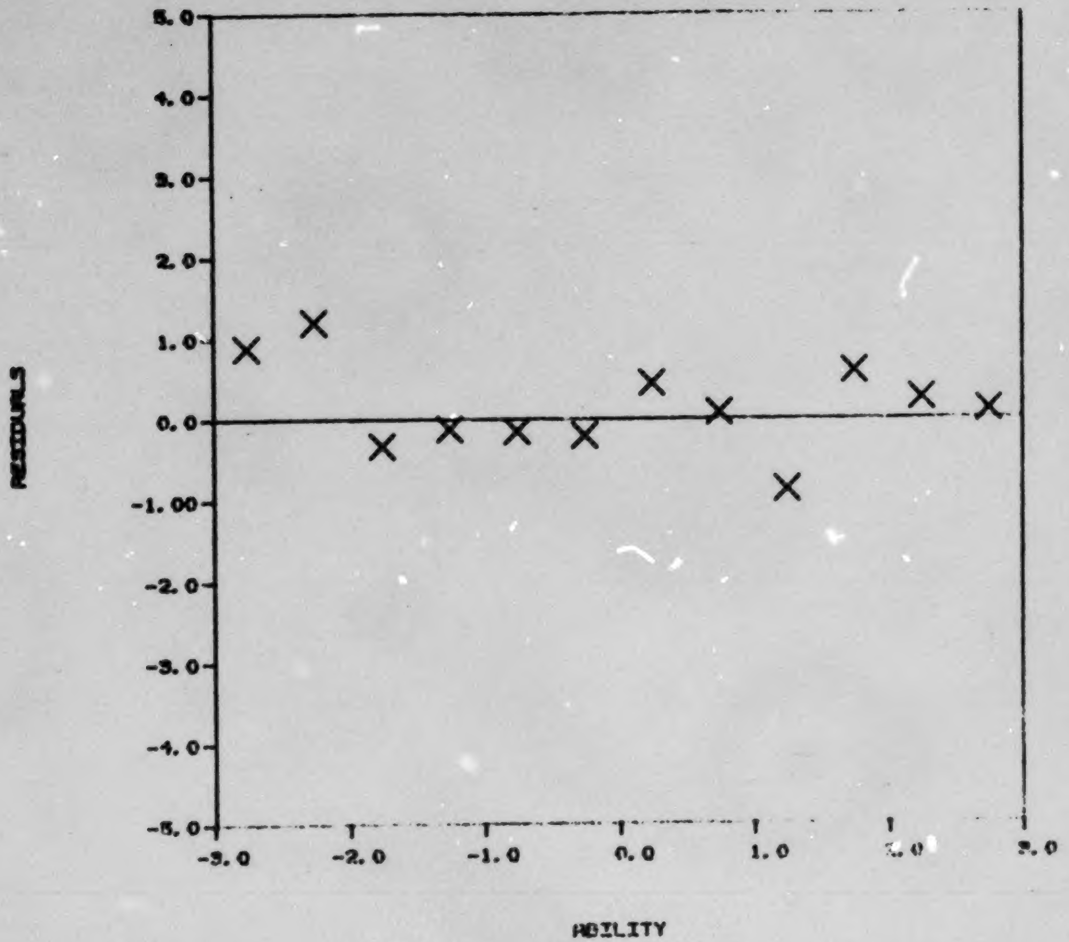


Figure 3.6.13 Standardized residual plot obtained with the three-parameter model for test item 2 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

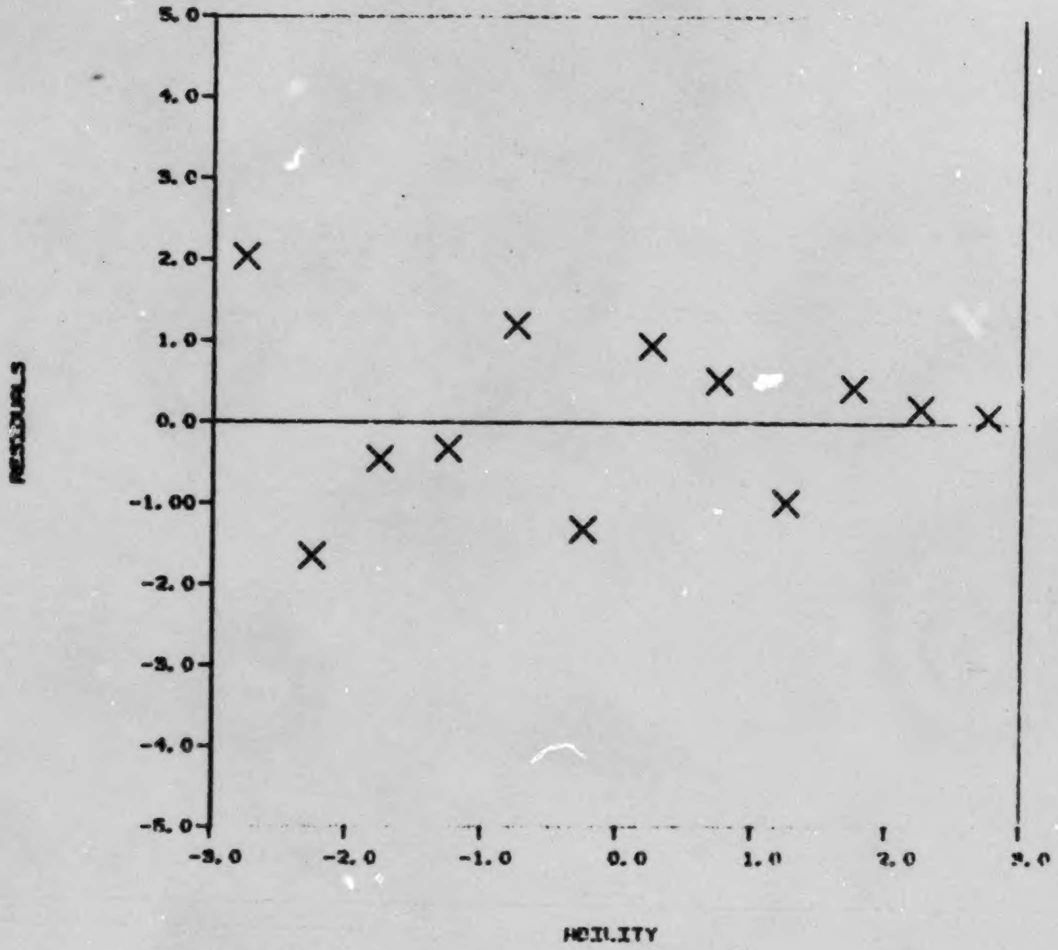


Figure 3.6.14 Standardized residual plot obtained with the three-parameter model for test item 3 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

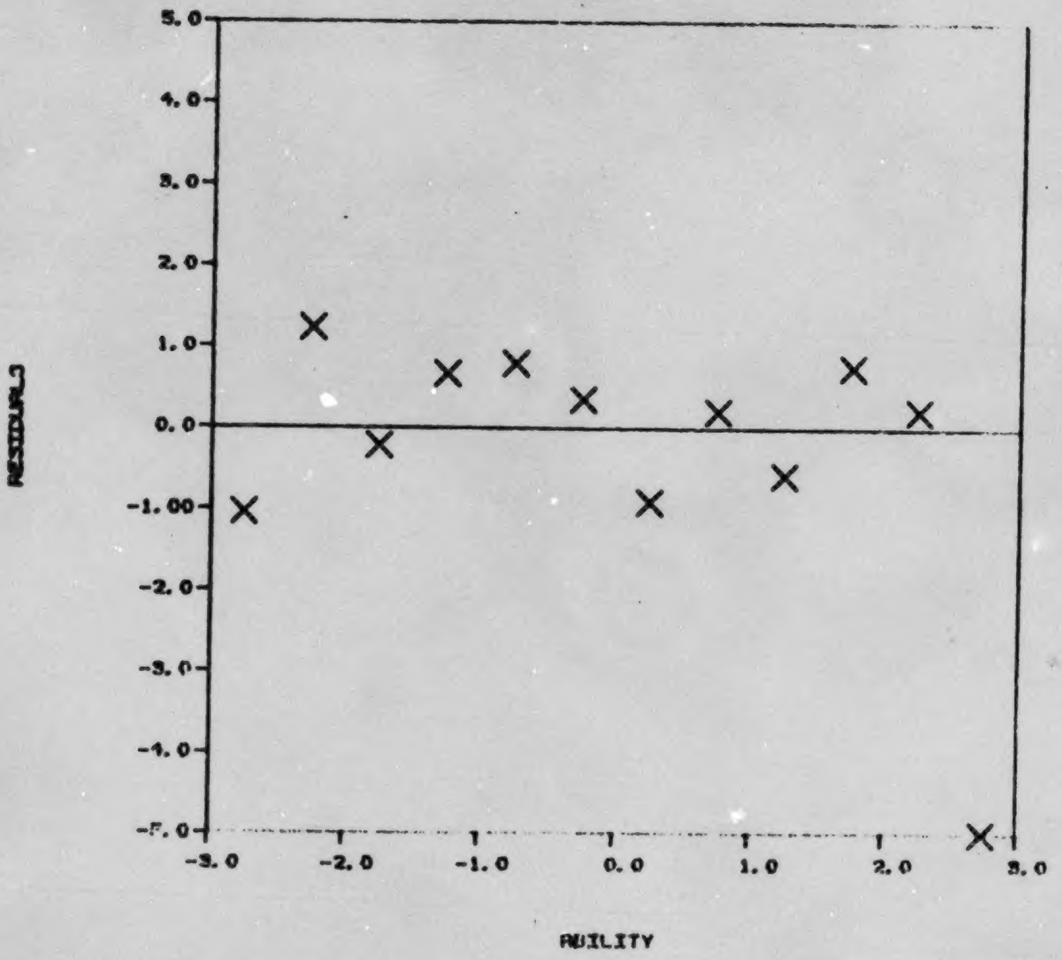


Figure 3.6.15 Standardized residual plot obtained with the three-parameter model for test item 4 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

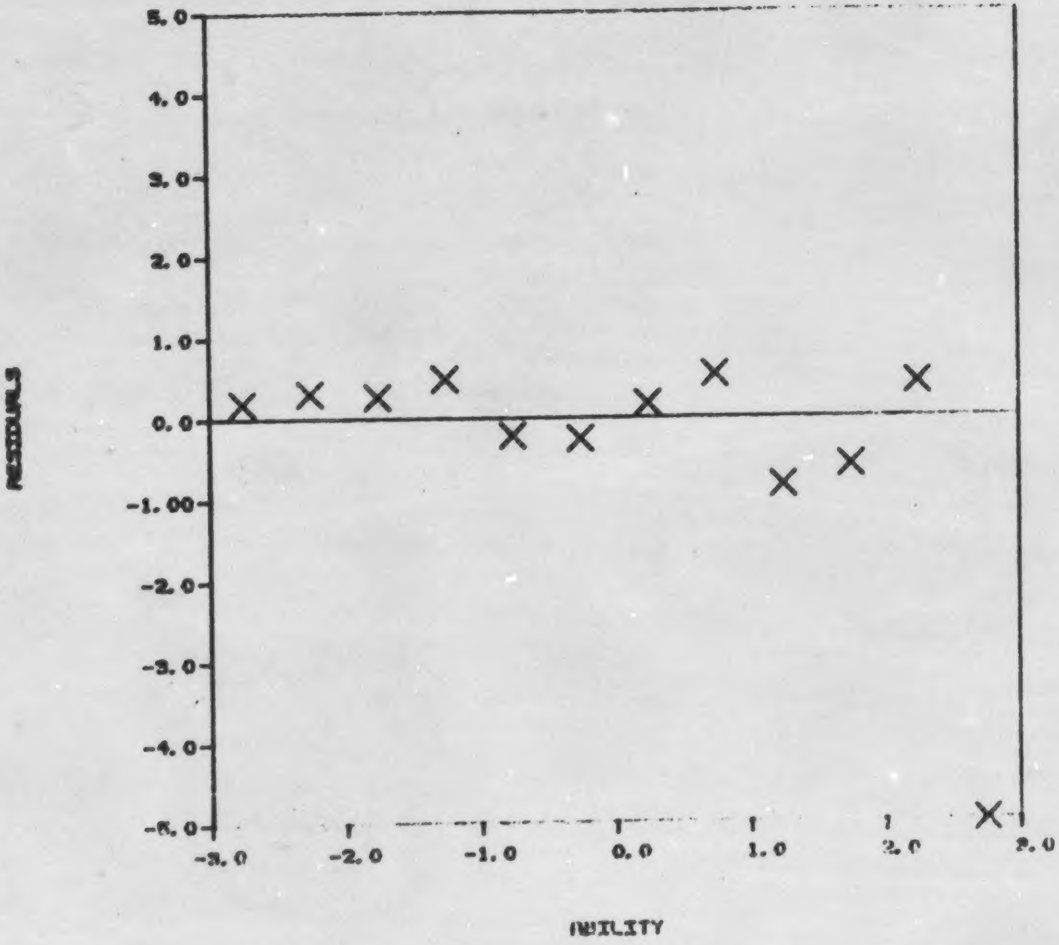


Figure 3.6.16 Standardized residual plot obtained with the three-parameter model for test item 5 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

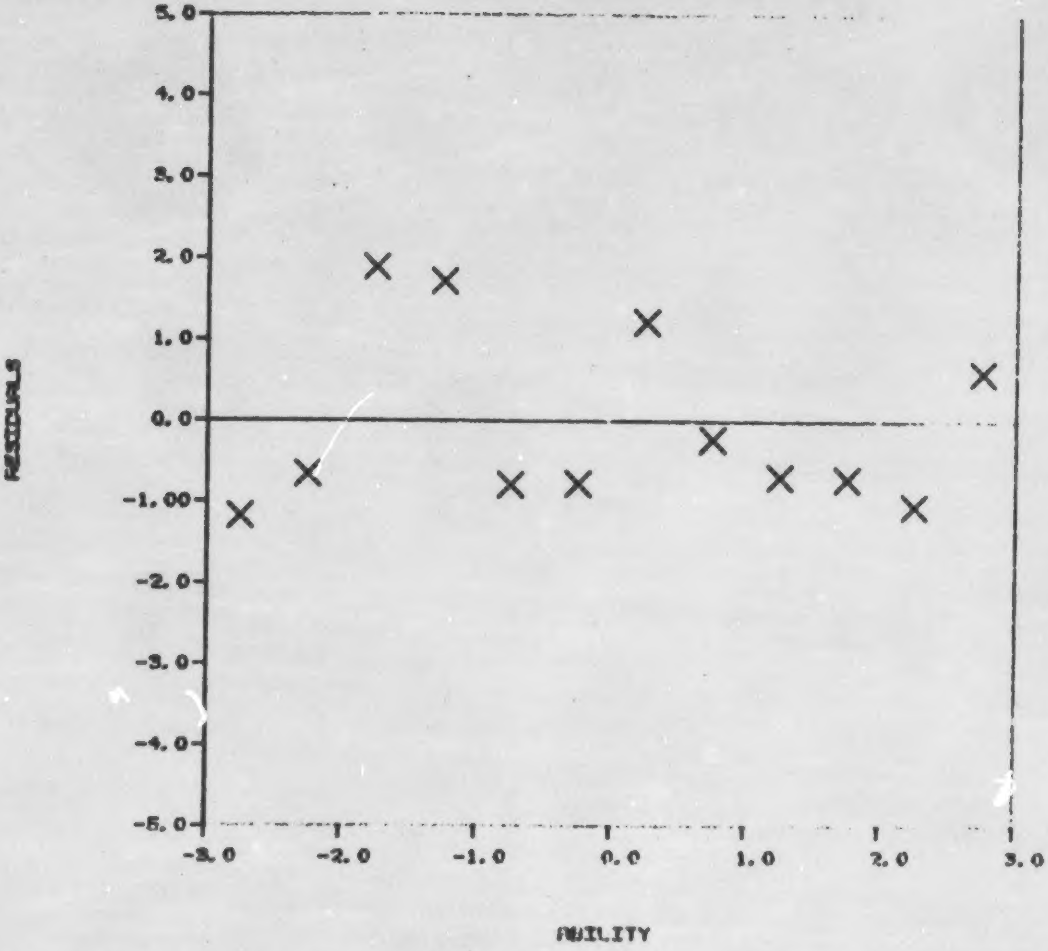


Figure 3.6.17 Standardized residual plot obtained with the three-parameter model for test item 6 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

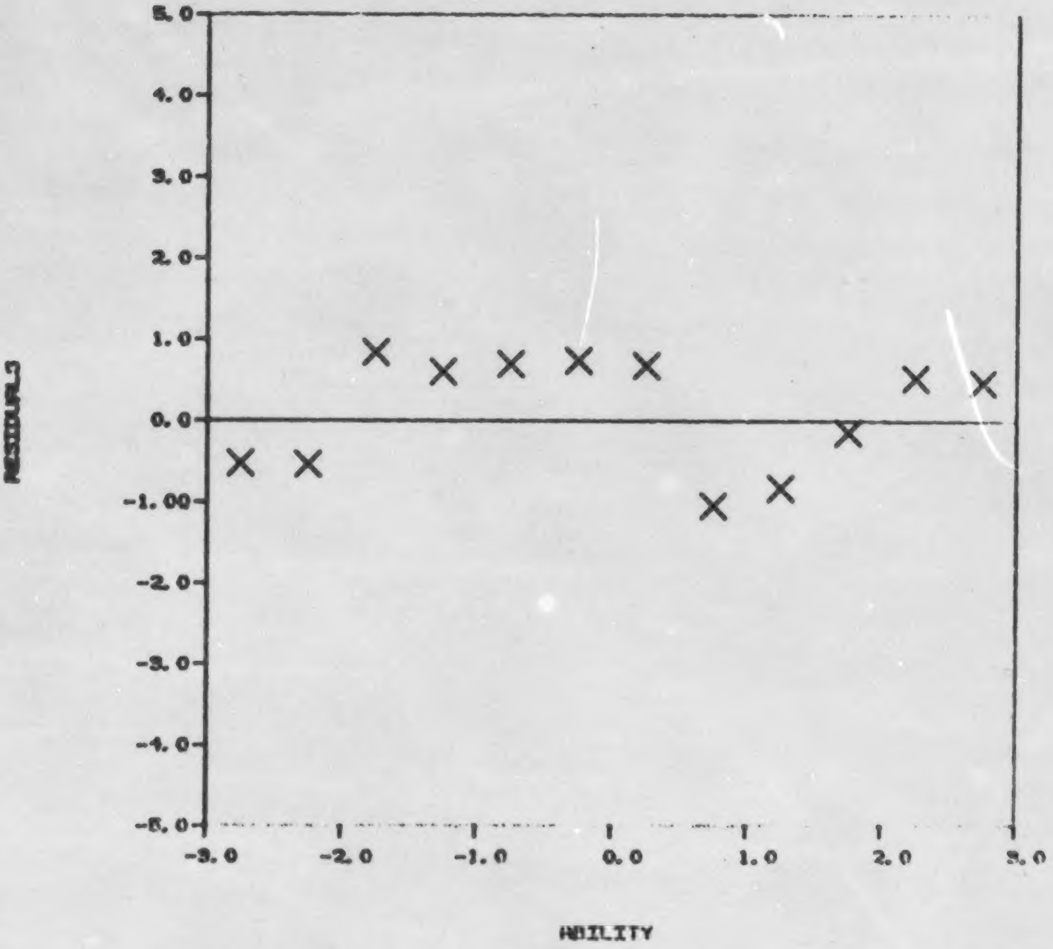


Figure 3.6.18 Standardized residual plot obtained with the three-parameter model for test item 7 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

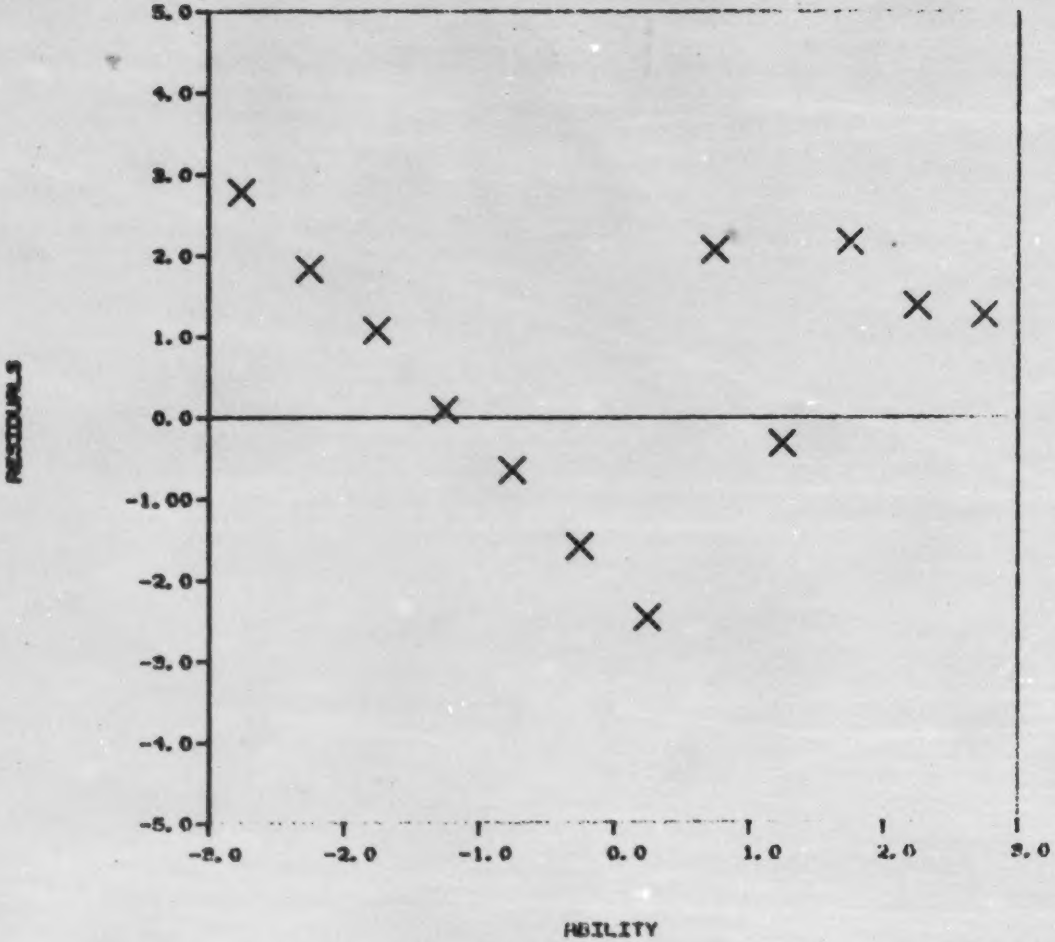


Figure 3.6.19 Standardized residual plot obtained with the three-parameter model for test item 8 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

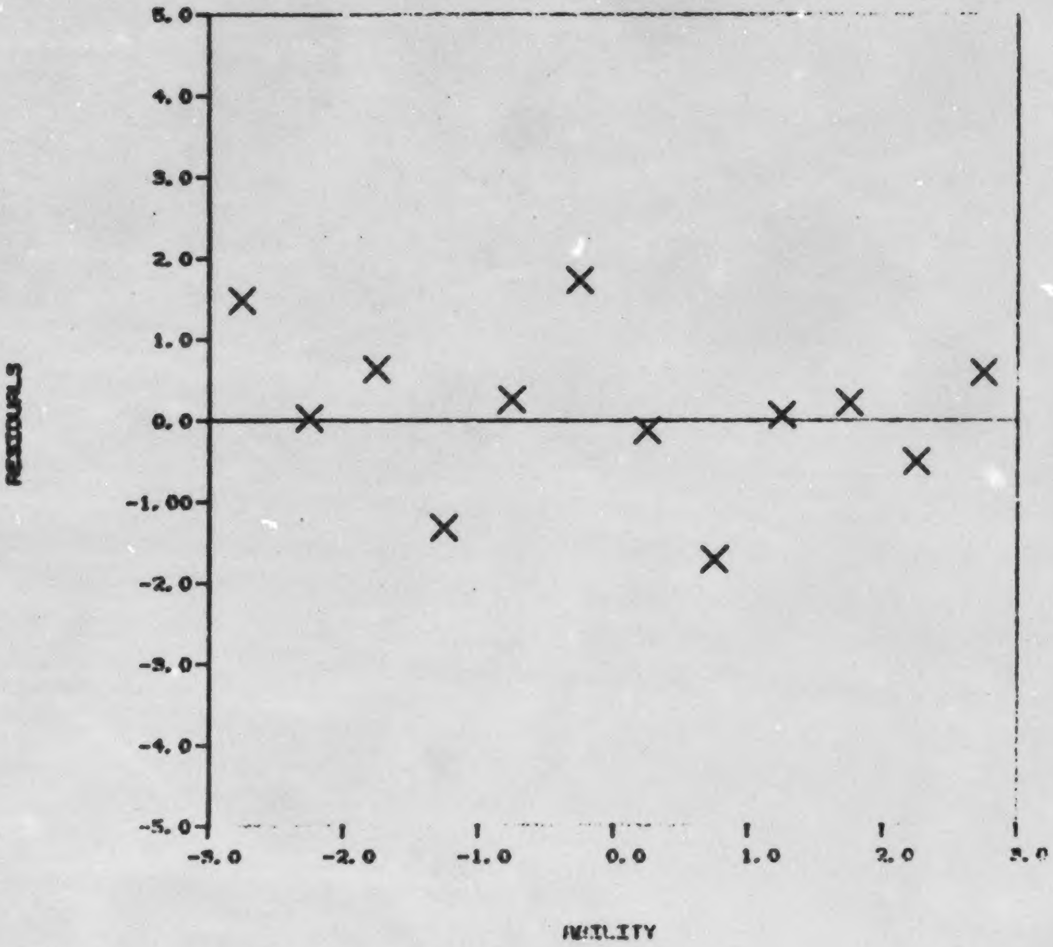


Figure 3.6.20 Standardized residual plot obtained with the three-parameter model for test item 9 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

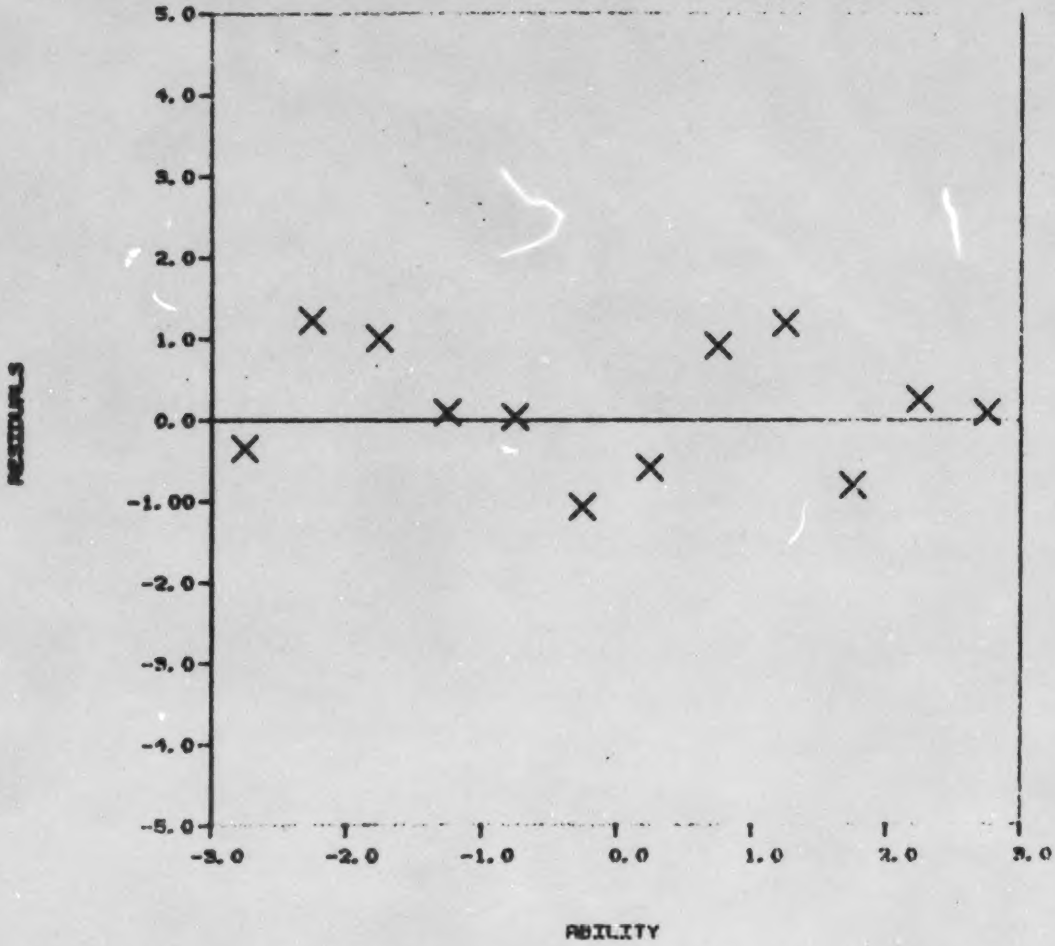


Figure 3.6.21 Standardized residual plot obtained with the three-parameter model for test item 10 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

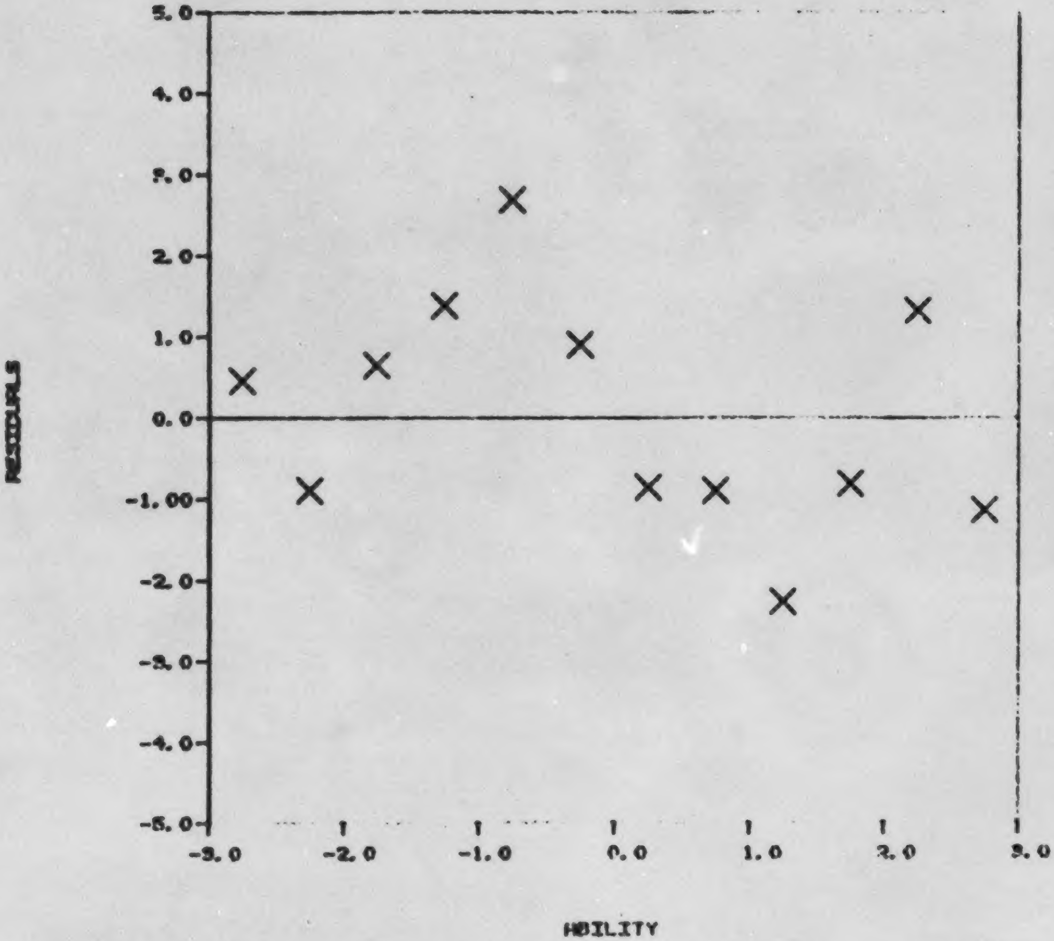


Figure 3.6.22 Standardized residual plot obtained with the three-parameter model for test item 36 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

Table 3.6.1

Analysis of Standardized Residuals with the One- and Three-Parameter Logistic Models for Six 1977-78 NAEP Mathematics Booklets

NAEP Booklet	Logistic Model	Percent of Residuals[1] |0 to 1|	|1 to 2|	|2 to 3|	|over 3|
Booklet 1	1	35.9	21.5	17.3	25.3
(9 Year Olds)	3	66.7	24.4	6.7	2.3
Booklet 2	1	37.1	25.3	13.8	23.8
(9 Year Olds)	3	67.4	24.7	5.7	2.2
Booklet 3	1	40.1	23.4	15.4	21.1
(9 Year Olds)	3	64.0	24.8	8.0	3.3
Booklet 1	1	40.7	22.1	16.5	20.7
(13 Year Olds)	3	65.4	25.1	7.8	1.7
Booklet 2	1	42.6	24.2	16.3	16.9
(13 Year Olds)	3	67.2	26.1	5.7	1.1
Booklet 3	1	34.3	24.5	17.5	23.7
(13 Year Olds)	3	61.0	26.8	8.2	4.0

---

[1] At the 9 Year Old Level, there were 780 standardized residuals (65 test items x 12 ability levels). At the 13 Year Old Level, there were 690 standardized residuals (58 test items x 12 ability levels).

Table 3.6.2

Analysis of Standardized Residuals at Twelve Ability Levels with the One- and Three-Parameter Logistic Models for Six 1977-78 NAEP Mathematics Booklets

NAEP Booklet	Test Length	Statistic	Logistic Model	Sample Size	Ability Level												Total (unweighted)
					-2.75	-2.25	-1.75	-1.25	-.75	-.25	.25	.75	1.25	1.75	2.25	2.75	
Booklet 1 (9 year olds)	65		1	2495	27	43	111	220	331	485	446	395	276	122	21	8	
			3	2495	29	50	108	212	333	454	470	403	271	100	21	9	
		Average Residual	1		.77	.99	.89	.79	.37	.20	.14	-.28	-.26	-.39	-.11	-.10	.25
			3		.00	.24	.27	.12	.16	.04	.08	-.18	-.48	-.36	-.32	-.16	.05
		Average Absolute Residual	1		1.75	2.40	2.82	3.35	2.35	1.80	1.62	2.35	2.64	2.40	1.19	.85	2.13
			3		.81	.90	1.02	.74	1.00	.94	.62	.87	.99	.85	.91	.88	.88
Booklet 2 (9 year olds)	75		1	2463	10	46	116	234	334	437	474	397	272	87	39	7	
			3	2463	23	64	89	218	346	417	497	403	230	107	34	6	
		Average Residual	1		.60	.74	.58	.71	.28	-.01	.02	-.14	-.02	.08	-.05	.02	.23
			3		-.16	-.14	-.02	.34	.50	.19	-.03	-.18	-.23	-.05	-.16	.01	.01
		Average Absolute Residual	1		1.55	2.42	3.02	3.10	2.28	1.49	1.59	2.36	2.75	1.71	1.31	.75	2.03
			3		.84	.95	.83	1.05	1.02	.94	1.04	.89	1.01	.87	.90	.53	.90

Table 3.6.2 (continued)

NAEP Booklet	Test Length	Statistic	Logistic Model	Sample Size	Ability Level												Total (unweighted)
					-2.75	-2.25	-1.75	-1.25	-.75	-.25	.25	.75	1.25	1.75	2.25	2.75	
Booklet 3	68		1	2438	29	44	120	174	326	410	533	446	186	89	44	9	
(9 year olds)			3	2438	28	62	108	177	283	438	319	410	219	88	29	9	
		Average Residual	1		.84	.57	.60	.17	-.01	-.04	.04	.30	.09	-.09	-.22	-.03	.19
			3		.36	.36	-.02	-.22	-.55	-.50	.01	.37	.39	.34	.09	.09	.06
		Average Absolute Residual	1		2.20	2.13	3.04	2.69	2.07	1.25	2.06	2.61	1.97	1.49	1.23	.95	1.97
			3		.94	.91	1.01	.97	1.03	.90	.97	.95	.90	1.18	.95	.64	.95
Booklet 1	58		1	2422	14	54	91	224	325	503	467	339	245	102	44	3	
(13 year olds)			3	2422	24	50	114	194	318	440	509	368	248	90	32	11	
		Average Residual	1		.67	.88	.66	.33	.03	.25	.40	.17	-.12	.07	.11	-.09	.28
			3		-.02	.08	.06	-.07	.07	.27	-.20	-.22	-.59	-.03	-.12	-.43	-.10
		Average Absolute Residual	1		1.76	2.59	2.74	2.64	2.20	1.63	1.68	2.08	2.06	1.62	1.30	.67	1.92
			3		1.27	1.02	.97	.84	.76	.92	.79	.84	1.07	.84	.86	.99	.93

Table 3.6.2 (continued)

NAEP Booklet	Test Length	Statistic	Logistic Model	Sample Size	Ability Level												Total (unweighted)
					-2.75	-2.25	-1.75	-1.25	-.75	-.25	.25	.75	1.25	1.75	2.25	2.75	
Booklet 2 (13 year olds)	62		1	2433	20	39	118	241	308	447	463	392	240	121	22	10	
			3	2433	15	45	121	230	334	429	440	402	259	94	26	11	
		Average Residual	1		.90	.92	1.06	.67	.14	-.05	-.01	-.05	-.27	-.13	-.03	-.01	.26
			3		-.03	.17	.13	.22	.03	.10	-.08	-.24	-.26	-.27	-.12	-.13	-.03
		Average Absolute Residual	1		2.08	1.90	2.98	2.69	1.64	1.79	1.54	1.95	2.10	1.64	1.00	.73	1.84
			3		.76	.79	.87	.82	.81	1.05	.94	.85	.85	.94	.69	.70	.84
Booklet 3 (13 year olds)	73		1	2469	12	38	96	215	400	540	403	341	237	120	51	8	
			3	2469	24	53	106	203	328	462	499	393	212	106	38	8	
		Average Residual	1		.45	.74	.78	.56	.28	.31	.23	-.16	-.01	.15	-.01	.10	.29
			3		.49	.08	.04	.27	.23	.01	-.15	-.15	-.34	-.31	-.36	-.13	-.03
		Average Absolute Residual	1		1.50	2.42	2.97	3.34	2.78	2.01	1.57	2.11	2.47	2.32	1.97	.84	2.19
			3		1.25	.94	1.29	1.05	.87	1.03	.97	.84	1.04	1.06	.99	.68	1.00

Table 3.6.3

NAEP Math Booklet No. 1
Basic Item Statistical and Classificatory Information
(9 year olds, 1977-78)

Test Item	Standardized Residuals[1]		Item Difficulty[2]	Item Discrimination[3]	Content Category[4]	Format[5]
	1-p	3-p				
1	1.27	0.62	.55	.62	3	1
2	1.73	0.60	.47	.69	3	1
3	1.27	0.85	.55	.65	1	1
4	3.50	2.24	.91	.34	2	1
5	2.28	1.57	.89	.39	2	1
6	3.26	1.08	.70	.33	2	1
7	2.00	0.88	.12	.37	2	2
8	0.59	0.82	.33	.56	2	2
9	1.73	0.63	.46	.47	2	1
10	1.53	0.63	.39	.65	5	1
11	2.18	0.79	.89	.77	4	2
12	2.03	1.01	.84	.75	4	2
13	2.45	0.84	.88	.80	4	2
14	2.35	1.73	.73	.76	4	2
15	2.61	1.06	.81	.80	4	2
16	3.05	2.16	.75	.79	4	2
17	3.20	1.00	.46	.35	1	1
18	0.49	0.59	.81	.59	1	2
19	0.86	1.30	.85	.51	2	1
20	0.85	0.73	.63	.63	4	2
21	2.35	0.48	.40	.75	4	2
22	2.26	0.74	.20	.60	5	1
23	1.84	0.65	.53	.62	1	1
24	2.50	0.58	.82	.79	4	2
25	1.55	0.86	.40	.68	4	2

[1] 1-p ≡ one-parameter logistic model; 3-p ≡ three-parameter logistic model.

[2] Item difficulty ≡ proportion of examinees in the NAEP sample answering the test item correctly (N = 2495).

[3] Item discrimination ≡ biserial correlation between item and the total test score.

[4] Content Categories: 1 - Story Problems, 2 - Geometry, 3 - Definitions, 4 - Calculations, 5 - Measurement, 6 - Graphs and Figures.

[5] Format: 1 - multiple choice, 2 - open response.

Table 3.6.3 (continued)

NAEP Math Booklet No. 1
Basic Item Statistical and Classificatory Information
(9 year olds, 1977-78)

Test Item	Standardized Residuals[1]		Item Difficulty[2]	Item Discrimination[3]	Content Category[4]	Format[5]
	1-p	3-p				
26	2.64	0.88	.49	.77	4	2
27	1.85	0.86	.68	.71	4	2
28	1.08	0.94	.36	.63	4	2
29	1.41	0.40	.77	.69	4	2
30	2.67	0.88	.68	.78	4	2
31	1.92	0.99	.69	.72	6	1
32	4.48	1.33	.03	.14	3	1
33	4.92	0.69	.19	.14	3	1
34	1.12	0.92	.64	.54	5	1
35	0.92	1.13	.80	.62	6	1
36	1.41	1.10	.65	.67	4	2
37	1.25	0.56	.09	.60	4	2
38	1.33	0.84	.94	.43	3	1
39	3.53	0.72	.20	.26	1	1
40	4.00	0.58	.17	.22	4	1
41	2.26	1.12	.20	.73	1	2
42	0.69	0.38	.17	.57	4	2
43	1.22	0.58	.02	.61	4	2
44	1.10	1.10	.01	.59	4	2
45	3.55	0.87	.29	.28	5	2
46	1.72	0.60	.36	.51	4	1
47	2.63	1.11	.54	.40	5	1
48	1.18	0.61	.83	.67	6	1
49	2.36	0.93	.29	.50	6	1
50	4.38	0.47	.66	.27	1	1
51	4.18	0.69	.25	.21	1	1
52	5.51	0.88	.35	.19	2	1
53	3.19	0.66	.09	.22	2	1
54	2.67	0.97	.09	.31	2	1
55	0.58	0.65	.01	.49	6	2
56	1.43	0.68	.12	.64	1	2
57	1.51	1.16	.48	.53	2	1
58	1.11	0.91	.24	.53	2	1
59	2.32	0.44	.28	.48	2	1
60	0.99	0.76	.21	.51	1	2
61	1.54	0.92	.10	.53	5	2
62	1.46	1.47	.85	.60	3	2
63	1.53	1.17	.48	.67	4	2
64	1.16	0.53	.35	.49	2	1
65	3.71	0.94	.27	.24	3	1

Table 3.6.4

NAEP Math Booklet No. 2
Basic Item Statistical and Classificatory Information
(9 year olds, 1977-78)

Test Item	Standardized Residuals[1] 1-p	3-p	Item Difficulty[2]	Item Discrimination[3]	Content Category[4]	Format[5]
1	3.27	0.67	.77	.31	3	1
2	3.20	0.64	.78	.31	3	1
3	0.73	0.90	.92	.60	4	2
4	1.50	0.77	.87	.70	4	2
5	1.38	1.27	.88	.65	4	2
6	1.35	1.22	.78	.67	4	2
7	1.67	0.96	.86	.71	4	2
8	1.44	0.88	.82	.70	4	2
9	2.39	1.16	.59	.76	4	2
10	2.57	0.79	.60	.76	4	2
11	2.87	0.65	.50	.78	4	2
12	2.34	0.79	.50	.74	4	2
13	0.94	0.59	.08	.46	2	1
14	1.00	0.83	.37	.58	1	1
15	1.19	1.31	.73	.57	6	1
16	1.31	0.71	.57	.63	6	1
17	1.03	0.77	.74	.64	4	2
18	1.06	0.73	.73	.65	4	2
19	1.59	1.06	.56	.68	4	2
20	1.31	0.99	.14	.56	1	1
21	1.77	0.55	.63	.71	6	1
22	2.17	1.01	.57	.72	6	1
23	2.26	1.06	.39	.71	6	1
24	1.18	0.67	.96	.68	3	1
25	0.83	0.70	.96	.60	3	1

[1] 1-p ≡ one-parameter logistic model; 3-p ≡ three-parameter logistic model.

[2] Item difficulty ≡ proportion of examinees in the NAEP sample answering the test item correctly (N = 2463).

[3] Item discrimination ≡ biserial correlation between item and the total test score.

[4] Content Categories: 1 - Story Problems, 2 - Geometry, 3 - Definitions, 4 - Calculations, 5 - Measurement, 6 - Graphs and Figures.

[5] Format: 1 - multiple choice, 2 - open response.

Table 3.6.4 (continued)

NAEP Math Booklet No. 2
Basic Item Statistical and Classificatory Information
(9 year olds, 1977-78)

Test Item	Standardized Residuals[1] 1-p	3-p	Item Difficulty[2]	Item Discrimination[3]	Content Category[4]	Format[5]
26	1.10	0.69	.97	.68	3	1
27	0.67	0.69	.94	.52	3	1
28	0.74	0.84	.92	.56	3	1
29	4.80	0.70	.19	.18	5	1
30	2.87	0.77	.20	.32	5	1
31	1.03	0.91	.25	.60	4	2
32	1.67	0.96	.27	.66	1	2
33	1.87	1.03	.49	.69	3	2
34	1.83	1.09	.52	.69	3	2
35	1.66	1.13	.47	.67	3	2
36	3.16	0.82	.39	.34	2	1
37	0.63	0.69	.84	.60	2	1
38	1.20	0.61	.19	.47	2	1
39	4.43	1.18	.25	.21	1	1
40	1.72	0.94	.63	.70	4	2
41	2.29	0.66	.40	.73	4	2
42	2.58	0.74	.72	.78	4	2
43	2.98	1.09	.56	.81	4	2
44	2.58	0.65	.74	.79	4	2
45	2.40	0.73	.46	.75	4	2
46	2.44	0.88	.19	.37	2	1
47	1.51	0.81	.90	.42	1	2
48	1.09	0.54	.75	.66	3	2
49	1.11	1.23	.50	.63	3	2
50	0.60	0.75	.41	.55	3	1
51	3.39	0.83	.80	.27	5	1
52	2.29	0.76	.71	.76	3	1
53	1.96	0.45	.50	.64	3	1
54	2.67	1.43	.44	.45	3	1
55	3.89	0.64	.25	.25	1	1
56	2.25	0.89	.54	.43	1	1
57	2.61	0.52	.37	.41	1	1
58	0.67	0.56	.66	.60	1	1
59	1.14	0.80	.50	.61	1	1
60	1.40	1.25	.23	.52	4	1

Table 3.6.4 (continued)

NAEP Math Booklet No. 2
Basic Item Statistical and Classificatory Information
(9 year olds, 1977-78)

Test Item	Standardized 1-p	Residuals[1] 3-p	Item Difficulty[2]	Item Discrimination[3]	Content Category[4]	Format[5]
61	4.08	5.44	.88	.13	2	1
62	3.07	0.73	.44	.35	2	1
63	4.76	0.56	.21	.16	3	1
64	5.88	0.84	.14	.06	3	1
65	4.63	0.54	.25	.19	3	1
66	0.81	0.66	.12	.45	4	2
67	1.68	1.78	.26	.50	1	2
68	0.82	0.48	.01	.54	2	2
69	2.15	1.05	.49	.42	1	1
70	2.63	0.94	.08	.22	1	2
71	1.65	0.67	.06	.35	2	2
72	1.21	0.63	.04	.58	4	2
73	1.76	0.83	.34	.44	5	2
74	0.59	0.99	.39	.57	6	2
75	2.66	0.74	.34	.35	5	1

Table 3.6.5

NAEP Math Booklet No. 1
Basic Item Statistical and Classificatory Information
(13 year olds, 1977-78)

Test Item	Standardized Residuals[1] 1-p	3-p	Item Difficulty[2]	Item Discrimination[3]	Content Category[4]	Format[5]
1	1.47	.84	.85	.70	1	2
2	.68	.44	.93	.61	3	1
3	.71	.85	.95	.62	3	1
4	3.11	1.94	.52	.81	5	2
5	1.74	.89	.65	.72	4	1
6	1.80	.96	.36	.48	2	1
7	1.70	.64	.40	.49	2	1
8	3.80	1.47	.70	.29	2	1
9	2.13	.72	.30	.43	1	1
10	1.59	.64	.81	.72	5	1
11	1.47	.86	.95	.75	4	2
12	1.47	1.31	.94	.74	4	2
13	1.61	1.11	.93	.75	4	2
14	1.21	.77	.92	.70	4	2
15	.97	.88	.89	.66	4	2
16	1.11	1.39	.88	.58	4	2
17	1.86	.98	.73	.47	5	1
18	.96	.83	.14	.54	1	2
19	2.42	1.42	.62	.75	4	2
20	3.30	.42	.59	.84	4	2
21	3.08	.53	.56	.82	4	2
22	.68	.48	.93	.46	3	1
23	2.85	.71	.36	.38	3	1
24	1.88	.89	.33	.48	3	1
25	1.15	.98	.52	.64	1	2

[1] 1-p ≡ one-parameter logistic model; 3-p ≡ three-parameter logistic model.

[2] Item difficulty ≡ proportion of examinees in the NAEP sample answering the test item correctly (N = 2500).

[3] Item discrimination ≡ biserial correlation between item and the total test score.

[4] Content Categories:
1 ≡ Story Problems, 2 ≡ Geometry, 3 ≡ Definitions, 4 ≡ Calculations, 5 ≡ Measurement, 6 ≡ Graphs and Figures.

[5] Format:
1 ≡ multiple-choice, 2 ≡ open response.

Table 3.6.5 (continued)

NAEP Math Booklet No. 1
Basic Item Statistical and Classificatory Information
(13 year olds, 1977-78)

Test Item	Standardized Residuals[1]		Item Difficulty[2]	Item Discrimination[3]	Content Category[4]	Format[5]
	1-p	3-p				
26	2.32	.46	.73	.41	2	1
27	1.06	.81	.10	.51	2	1
28	4.62	.77	.22	.18	2	2
29	.92	.77	.18	.57	5	2
30	1.92	.83	.46	.60	1	1
31	.80	.73	.74	.64	2	1
32	2.06	1.56	.58	.64	1	1
33	1.13	.64	.42	.49	1	1
34	.75	.56	.96	.46	2	1
35	2.36	1.87	.66	.44	2	1
36	7.08	1.19	.21	-.01	1	1
37	1.36	.66	.37	.47	2	1
38	2.63	.67	.78	.80	3	1
39	3.37	.73	.70	.36	3	1
40	1.72	.85	.66	.70	1	1
41	1.16	.96	.27	.62	3	1
42	.60	.93	.69	.60	2	1
43	.87	.81	.78	.60	2	1
44	1.58	1.93	.68	.59	4	2
45	1.16	1.62	.45	.61	4	2
46	2.01	.90	.34	.63	1	1
47	4.63	.98	.11	.10	2	1
48	1.69	1.11	.15	.48	3	1
49	1.20	.83	.49	.64	4	2
50	.77	.80	.84	.62	1	1
51	3.30	.57	.18	.27	1	1
52	5.03	.96	.60	.26	1	1
53	1.37	.31	.82	.45	2	1
54	1.19	1.19	.73	.63	4	2
55	1.83	.83	.25	.68	6	2
56	.49	.74	.72	.59	1	1
57	2.48	.95	.31	.73	5	2
58	.83	.71	.74	.62	4	2

Table 3.6.6

NAEP Math Booklet No. 2
Basic Item Statistical and Classificatory Information
(13 year olds, 1977-78)

Test Item	Standardized Residuals[1] 1-p	3-p	Item Difficulty[2]	Item Discrimination[3]	Content Category[4]	Format[5]
1	1.01	1.06	.58	.60	4	2
2	1.13	0.85	.48	.67	4	2
3	2.39	1.74	.65	.53	4	2
4	1.92	0.72	.69	.50	1	1
5	1.49	0.86	.57	.69	3	1
6	0.87	1.03	.18	.55	2	2
7	1.00	1.15	.51	.63	5	2
8	0.56	0.53	.96	.58	1	2
9	2.25	0.52	.85	.84	4	2
10	2.33	0.62	.84	.84	4	2
11	2.20	1.31	.82	.84	4	2
12	2.11	0.56	.79	.82	4	2
13	0.93	0.67	.92	.68	2	1
14	2.17	0.92	.42	.48	4	1
15	1.20	1.02	.30	.61	2	1
16	0.71	0.61	.89	.66	4	2
17	0.79	0.55	.85	.69	4	2
18	0.93	0.51	.86	.70	4	2
19	1.00	0.77	.95	.50	4	2
20	0.99	0.94	.95	.68	4	2
21	1.13	0.76	.95	.56	4	2
22	6.17	1.14	.06	-.07	2	1
23	1.77	0.66	.38	.74	4	2
24	1.57	0.71	.45	.74	4	2
25	1.12	1.20	.61	.63	4	2

[1] 1-p ≡ one-parameter logistic model; 3-p ≡ three-parameter logistic model.

[2] Item difficulty ≡ proportion of examinees in the NAEP sample answering the test item correctly (N = 2433).

[3] Item discrimination ≡ biserial correlation between item and the total test score.

[4] Content Categories: 1 - Story Problems, 2 - Geometry, 3 - Definitions, 4 - Calculations, 5 - Measurement, 6 - Graphs and Figures.

[5] Format: 1 - multiple choice, 2 - open response.

Table 3.6.6 (continued)

NAEP Math Booklet No. 2
Basic Item Statistical and Classificatory Information
(13 year olds, 1977-78)

Item	Standardized Residuals[1]		Item Difficulty[2]	Item Discrimination[3]	Content Category[4]	Format[5]
	1-p	3-p				
26	3.45	1.00	.88	.24	2	1
27	3.63	0.89	.55	.36	2	1
28	3.24	1.48	.24	.49	1	2
29	0.62	0.90	.91	.59	1	1
30	1.07	1.25	.16	.54	1	2
31	1.54	0.67	.30	.67	3	1
32	3.03	0.99	.67	.44	3	1
33	1.05	0.33	.95	.77	3	1
34	0.74	0.62	.86	.65	1	1
35	1.02	1.16	.22	.57	1	2
36	0.74	0.55	.59	.64	6	1
37	2.20	0.65	.67	.77	6	1
38	1.53	0.70	.34	.61	6	1
39	0.62	0.60	.50	.64	4	2
40	1.46	0.76	.45	.64	1	1
41	0.85	0.70	.88	.69	4	2
42	1.80	1.69	.78	.73	4	2
43	0.81	0.79	.78	.59	1	1
44	3.61	0.80	.73	.37	2	1
45	1.64	0.76	.66	.53	1	1
46	1.08	0.77	.81	.68	1	1
47	1.36	0.62	.80	.76	1	1
48	1.24	0.84	.26	.65	3	2
49	1.83	0.36	.17	.68	3	2
50	1.51	1.06	.63	.72	3	2
51	6.21	1.28	.32	.17	2	1
52	2.99	0.65	.17	.32	2	1
53	2.13	0.51	.38	.75	1	2
54	1.23	0.69	.86	.55	2	1
55	1.05	0.53	.47	.56	2	1
56	2.41	0.89	.50	.80	1	2
57	6.38	0.76	.13	.10	1	1
58	2.53	0.78	.17	.56	6	1
59	3.57	1.19	.19	.45	6	1
60	1.12	0.64	.75	.56	6	1
61	1.06	0.92	.64	.58	4	2
62	1.71	0.83	.29	.70	4	2

Our initial preliminary studies are reported in Figures 3.6.23 to 3.6.28. Figure 3.6.23 shows the relationship between one-parameter model residuals and classical item difficulties. The outstanding features are the large size of the residuals and the tendency for the most difficult items to have the highest residuals. Possibly this latter problem is due to the guessing behavior of examinees. In a similar plot with three-parameter model residuals shown in Figure 3.6.24, the standardized residuals are substantially smaller and it appears that by estimating item pseudo-chance level parameters, the tendency for the highest residuals to be obtained with the most difficult items is reduced.

Figure 3.6.25 provides a plot of one-parameter model standardized residuals and classical item discrimination indices for four of the Math Booklets combined. A strong curvilinear relationship is evident. Items with relatively high or low classical discrimination indices have the highest standardized residuals. Figure 3.6.26 provides the same plot using three-parameter model standardized residuals. The curvilinear relationship disappears. Substantially better fits are obtained when variations in discriminating powers of test items are handled in the chosen model.

Figures 3.6.27 and 3.6.28 provide comparable information to the previous two figures except that the latter two figures use the information from a single test booklet. The trends in the results are identical.

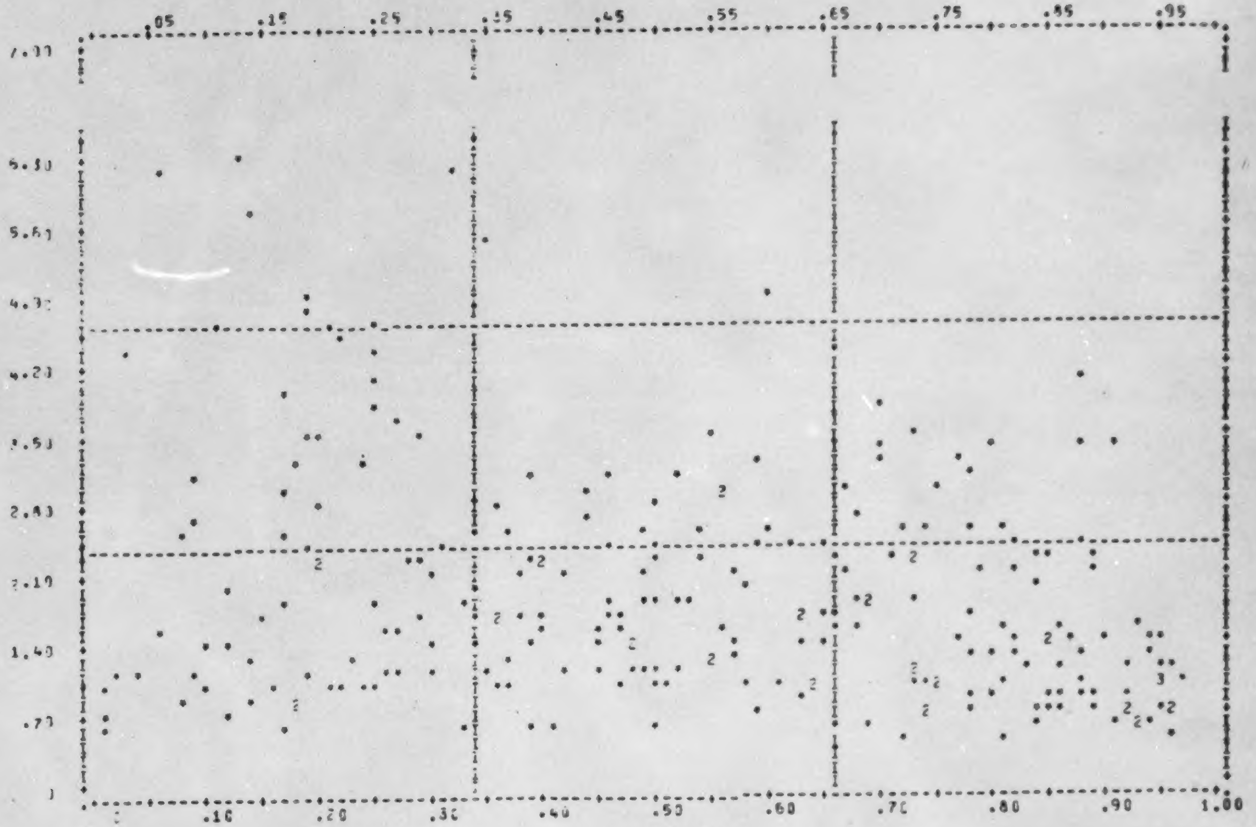


Figure 3.6.23. Scatterplot of one-parameter standardized residuals and item difficulties for 9 and 13 Year Old Math Booklets Nos. 1 and 2.

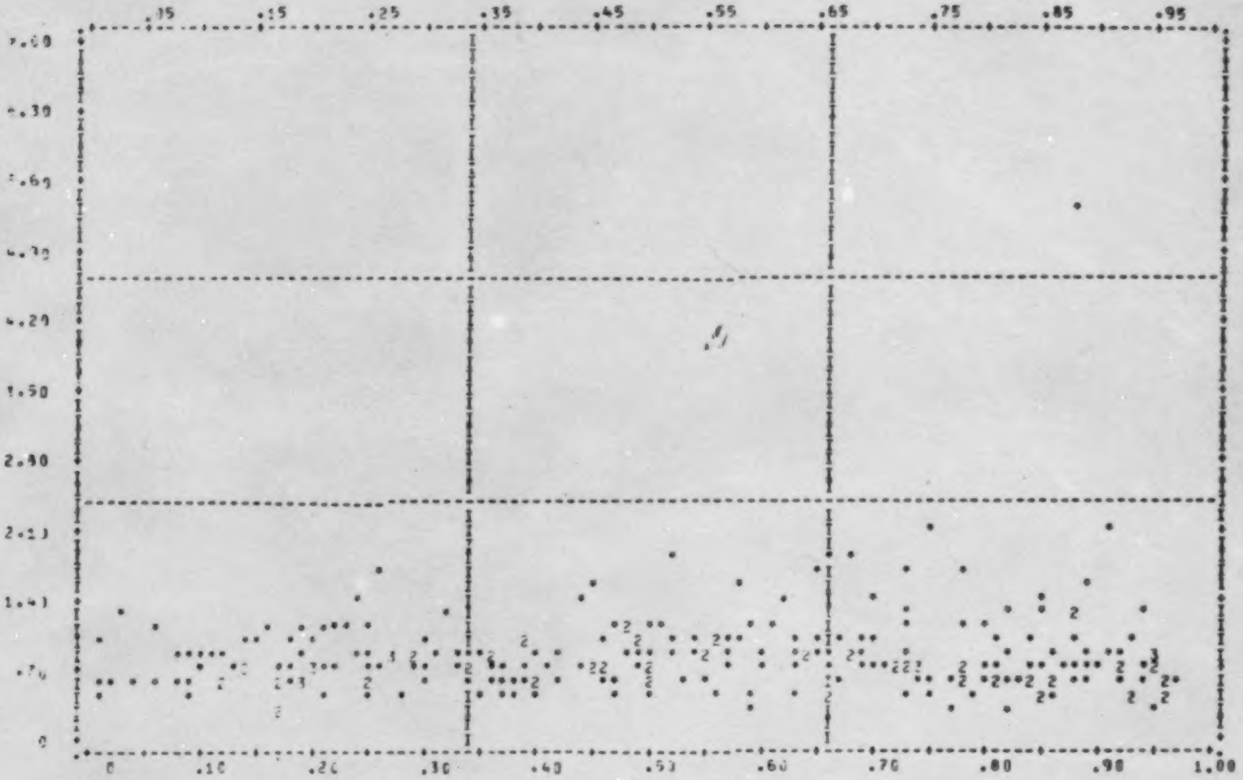


Figure 3.6.24. Scatterplot of three-parameter standardized residuals and item difficulties for 9 and 13 Year Old Math Booklets Nos. 1 and 2.

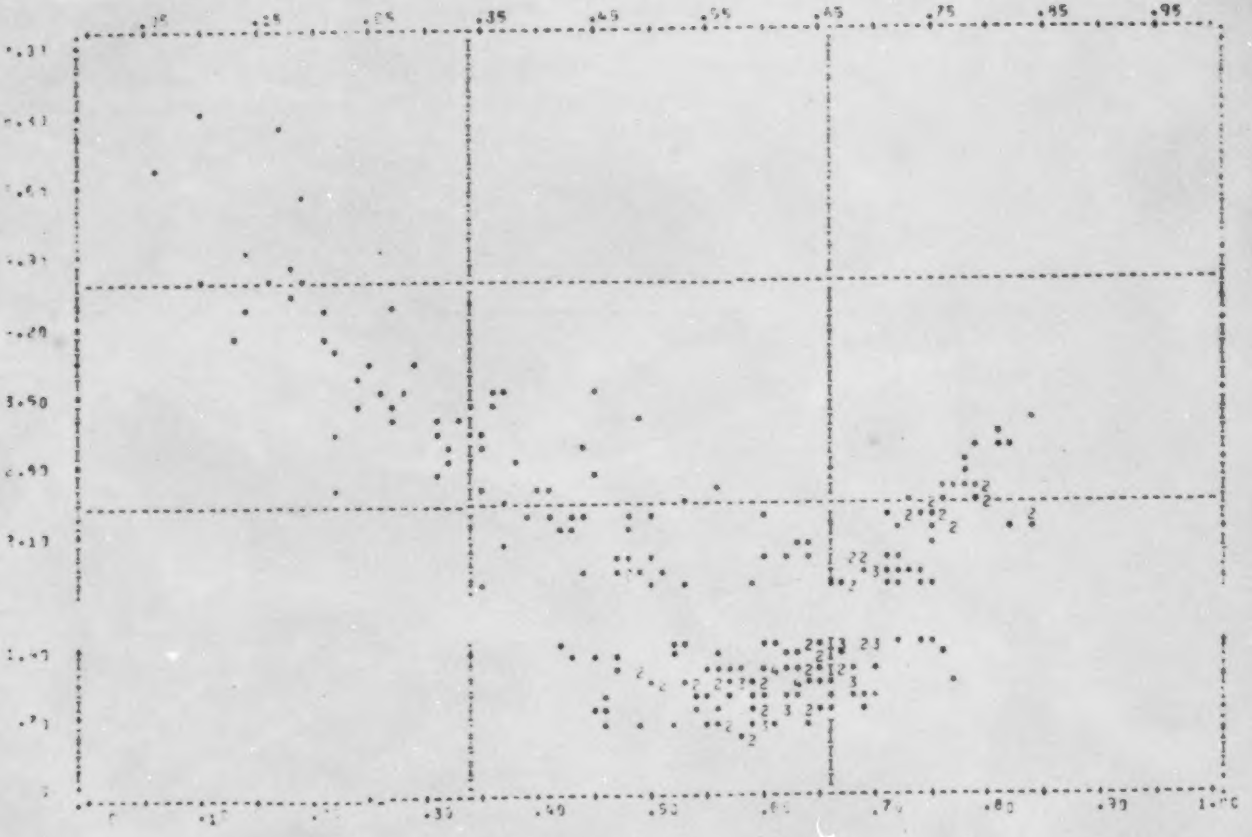


Figure 3.6.25. Scatterplot of one-parameter standardized residuals and item discrimination indices for 9 and 13 Year Old Math Booklets Nos. 1 and 2.

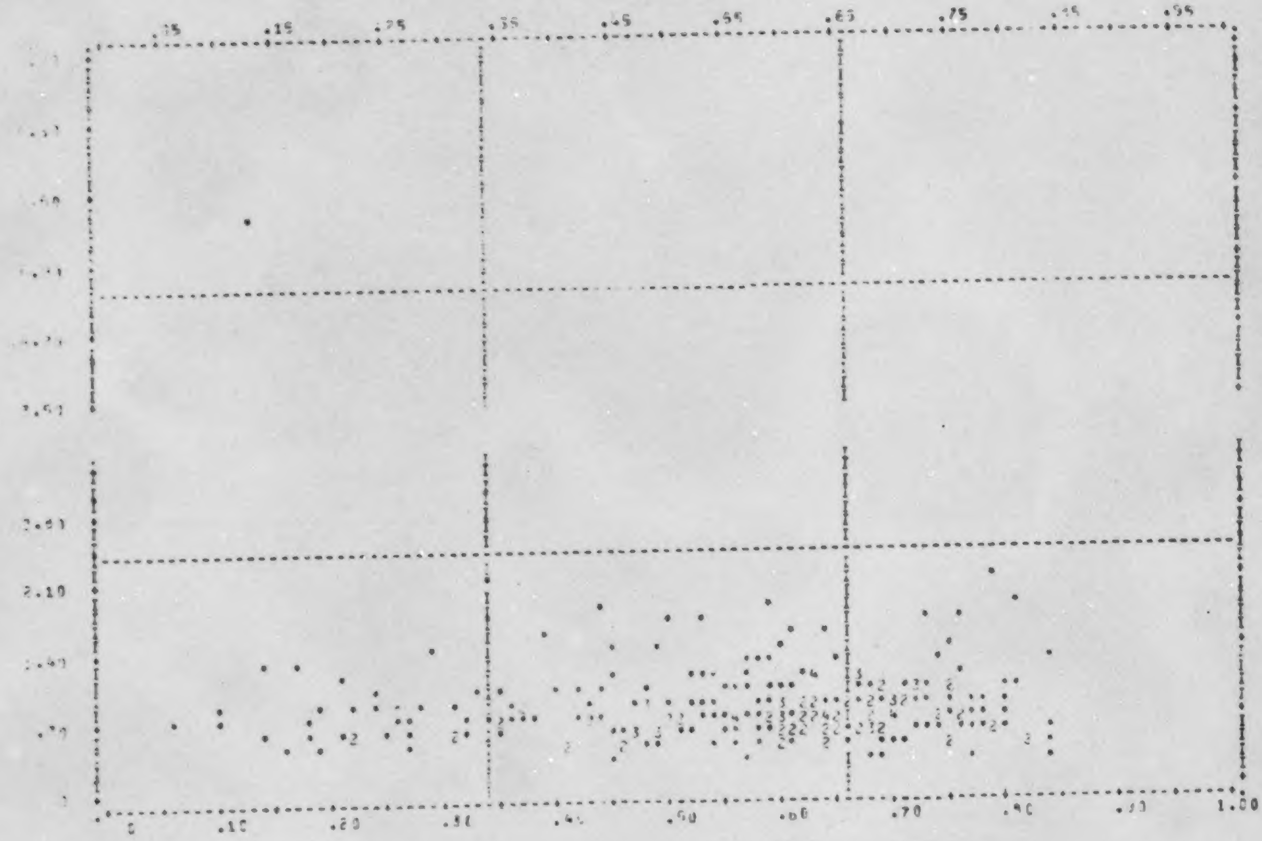


Figure 3.6.26. Scatterplot of three-parameter stardardized residuals and item discrimination indices for 9 and 13 Year Old Math Booklets Nos. 1 and 2.

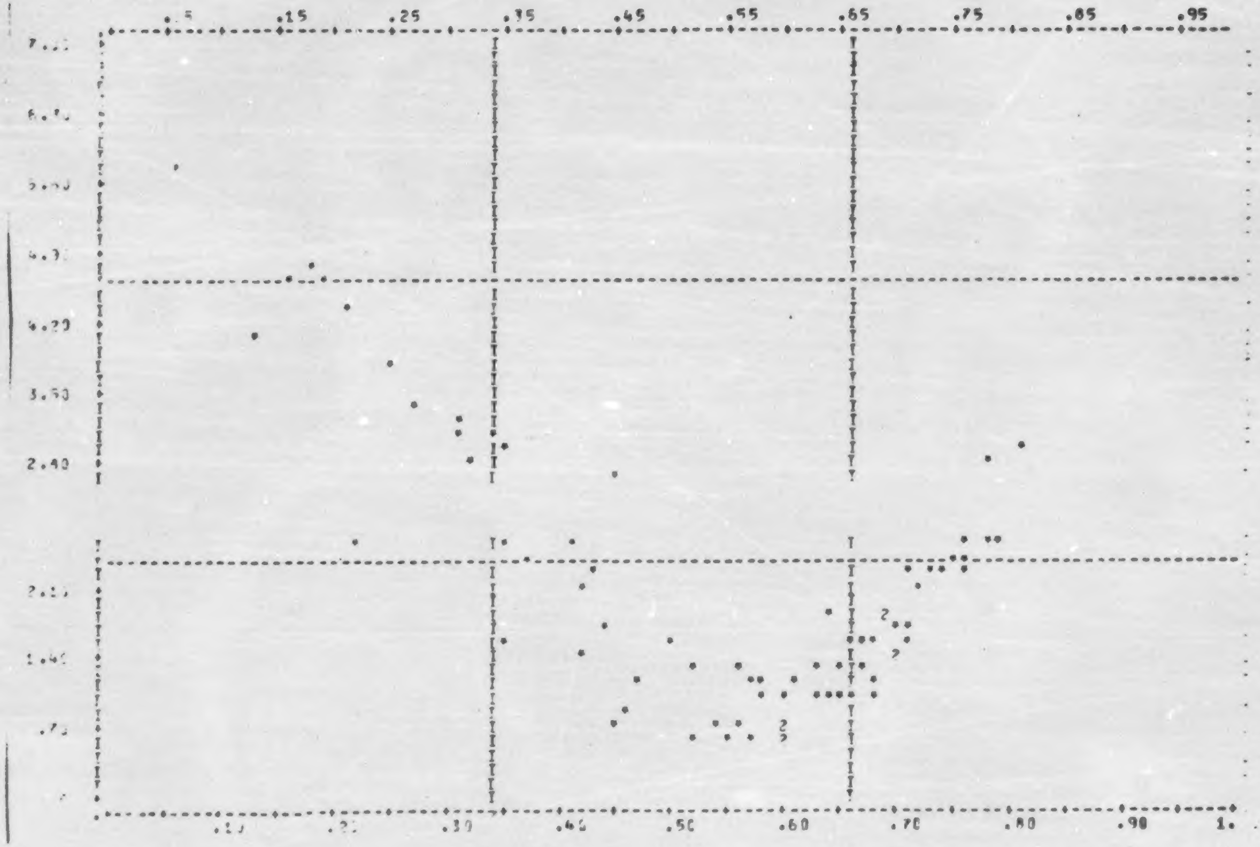


Figure 3.6.27. Scatterplot of one-parameter standardized residuals and item discrimination indices for 9 Year Old Math Booklet No. 2.

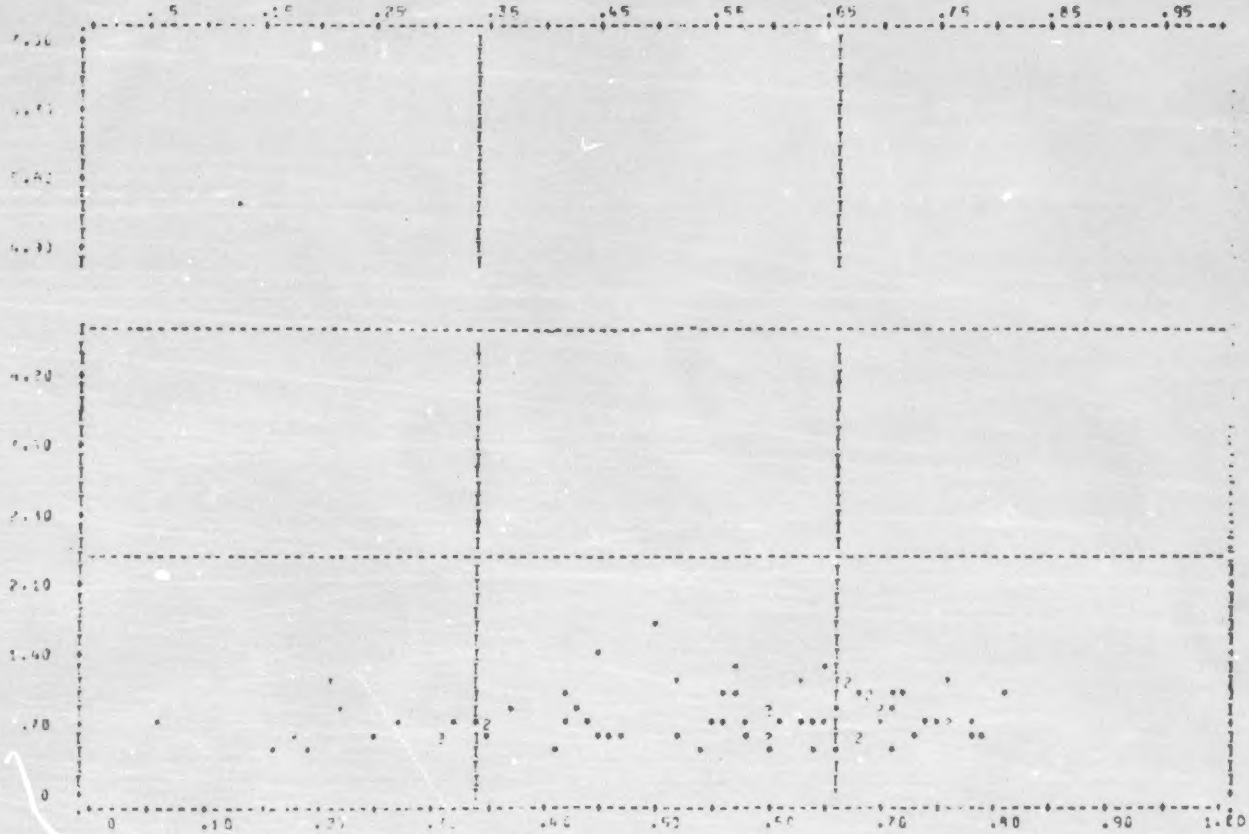


Figure 3.6.28. Scatterplot of three-parameter standardized residuals and item discrimination indices for 9 Year Old Math Booklet No. 2

These initial analyses were encouraging because they provided several insights into possible reasons for item misfit. Next, a more comprehensive analysis of the test items was initiated. Seven different analyses were carried out on four of the test booklets. In addition, the analyses were carried out on a combined set of Math Booklets.

Math Booklet	Tables
No. 1, 9 Year Olds	3.6.7 to 3.6.13
No. 2, 9 Year Olds	3.6.14 to 3.6.20
No. 1, 13 Year Olds	3.6.21 to 3.6.27
No. 2, 13 Year Olds	3.6.28 to 3.6.34
Combined	3.6.35 to 3.6.41

By combining booklets and obtaining more test items it was possible to more clearly study the trends in the results.

Since the trends in all of the analyses at the Math Booklet level are the same, only the results for the combined Math Booklets will be discussed further:

Table 3.6.35

- Intercorrelations among five key variables.

1. There is a high negative correlation (r=-.61) between one-parameter standardized residuals and classical item discrimination indices.[1] The result suggests that the poorest fitting items are the least discriminating. Perhaps this is due, in part, to examinee guessing behavior.

2. The most difficult test items are the least discriminating (r=.41). Again, perhaps the result is due to examinee guessing on hard test items.

3. There is a substantial correlation (r=.49) between item format and classical item discrimination indices. Open-ended test items tend to have higher discrimination indices than do multiple-choice items. Again, it is noted that guessing is a factor in multiple-choice test performance but plays almost no part with open-ended test items.

4. The higher one-parameter model residuals are associated with the multiple-choice test items; the lower one-parameter model residuals are associated with the open-ended items.

---

[1] This correlation is misleading because the actual relationship between the two variables is non-linear.

Table 3.6.7

Correlations Among Several NAEP Math Item Variables
(Booklet No. 1, 65 Items, 9 Year Olds, 1977-78)

Variable	SR(1-p)	SR(3-p)	p	r	F[1]
Item Order	.05	-.19	-.46	-.32	-.06
Standardized Residual (1-p)		.16	-.11	-.60	-.34
Standardized Residual (3-p)			.35	.03	.02
Item Difficulty (p)				.43	.02
Item Discrimination (r)					.57
Format (F)					

[1] 1=Multiple-Choice; 2=Open-Ended.

Table 3.6.8

Association Between Standardized Residuals and NAEP Item Content Classifications (Booklet No. 1, 65 Items, 9 Year Olds, 1977-78)

		Standardized Residuals			
		1-p		3-p	
Content Category	Number of Items	SR($\leq$1.0) (n= 8)	SR(>1.0) (n= 57)	SR($\leq$1.0) (n= 48)	SR(>1.0) (n= 17)
Story Problems	10	20.0	80.0	90.0	10.0
Geometry	14	14.3	85.7	64.3	35.7
Definitions	7	0.0	100.0	71.4	28.6
Calculations	23	8.7	91.3	69.6	30.4
Measurement	6	0.0	100.0	83.3	16.7
Graphs and Figures	5	40.0	60.0	80.0	20.0
		χ^2= 6.25		χ^2= 2.63	
		d.f.= 5	p= .282	d.f.=5	p= .757

Table 3.6.9

Association Between Standardized Residuals and Item Formats
(Booklet No. 1, 65 Items, 9 Year Olds, 1977-78)

Format	Standardized Residuals	1-p Results N	1-p Results %	3-p Results N	3-p Results %
Multiple-Choice	SR($\leq$1.0)	2	3.1	26	40.0
	SR(>1.0)	32	49.2	8	12.3
Open-Ended	SR($\leq$1.0)	6	9.2	22	33.8
	SR(>1.0)	25	38.5	9	13.6
		χ^2= 1.62		χ^2= .049	
		d.f.= 1	p= .203	d.f.= 1	p= .825

Table 3.6.10

Association Between Standardized Residuals and Item Difficulties
(Booklet No. 1, 65 Items, 9 Year Olds, 1977-78)

Difficulty Level	Standardized Residuals	1-p Results N	1-p Results %	3-p Results N	3-p Results %
Hard ($p<.5$)	SR($\leq$1.0)	4	6.2	32	49.2
	SR(>1.0)	33	50.8	5	7.7
Easy ($p\geq.5$)	SR($\leq$1.0)	4	6.2	16	24.6
	SR(>1.0)	24	36.9	12	18.5
		$\chi^2 = .102$		$\chi^2 = 5.66$	
		d.f.= 1	p= .967	d.f.= 1	p= .017

Table 3.6.11

Association Between Item Formats and Item Difficulties
(Booklet No. 1, 65 Items, 9 Year Olds, 1977-78)

Difficulty Level	Format	N	%
Hard ($p<.5$)	Multiple-Choice	20	30.8
	Open-Ended	17	26.2
Easy ($p\geq.5$)	Multiple-Choice	14	21.5
	Open-Ended	14	21.5

$\chi^2 = .005$

d.f.= 1 p= .942

Table 3.6.12

Descriptive Statistical Analysis of Standardized Residuals
(Booklet No. 1, 65 Items, 9 Year Olds, 1977-78)

Difficulty Level	Format	Number of Items	1-p Results		3-p Results	
			$\overline{X}$	SD	$\overline{X}$	SD
Hard ($p<.5$)	Multiple-Choice	20	2.84	1.31	.78	.23
	Open-Ended	17	1.55	.79	.80	.23
Easy ($p\geq.5$)	Multiple-Choice	14	1.98	1.09	1.03	.46
	Open-Ended	14	1.95	.74	1.01	.48

Table 3.6.13

Relationship Between Item Discrimination Indices and Standardized Residuals

(Booklet No. 1, 65 Items, 9 Year Olds, 1977-78)

Model	Standardized Residuals	-.01 to .30	.31 to .50	.51 to .70	.71 to 1.00
		(10)[1]	(13)	(29)	(13)
1-p	0.00 to 1.00	0.0	7.7	24.1	0.0
	1.01 to 2.00	0.0	30.8	72.4	15.4
	over 2.00	100.0	61.5	3.4	84.6
		χ^2= 42.24 Eta=.743	d.f.= 6	p=.010	
3-p	0.00 to 1.00	90.0	69.2	75.9	61.5
	1.01 to 2.00	10.0	23.1	24.1	30.8
	over 2.00	0.0	7.7	0.0	7.6
		χ^2=4.76 Eta=.231	d.f.= 6	p=.575	

Discrimination Indices is the spanning header over the last four columns.

---

[1] Number of test items appear in brackets.

Table 3.6.14

Correlations Among Several NAEP Math Item Variables (Booklet No. 2, 75 Items, 9 year Olds, 1977-78)

Variable	SR(1-p)	SR(3-p)	p	r	F[1]
Item Order	.19	.08	-.51	-.44	-.09
Standardized Residual (1-p)		.17	-.26	-.60	-.27
Standardized Residual (3-p)			.15	-.19	-.01
Item Difficulty (p)				.40	.03
Item Discrimination (r)					.49
Format (F)					

[1] 1=Multiple-Choice; 2=Open-Ended.

Table 3.6.15

Association Between Standardized Residuals
and NAEP Item Content Classifications
(Booklet No. 2, 75 Items, 9 Year Olds, 1977-78)

		Standardized Residuals			
		1-p		3-p	
Content Category	Number of Items	SR(≤1.0) (n= 12)	SR(>1.0) (n= 63)	SR(≤1.0) (n= 57)	SR(>1.0) (n= 18)
Story Problems	13	15.4	84.6	76.9	23.1
Geometry	9	33.3	66.7	88.9	11.1
Definitions	19	21.1	78.9	73.7	26.3
Calculations	23	8.7	91.3	73.9	26.1
Measurement	5	0.0	100.0	100.0	0.0
Graphs and Figures	6	16.7	83.3	50.0	50.0
		$\chi^2 = 4.24$		$\chi^2 = 4.74$	
		d.f.= 5	p= .515	d.f.= 5	p= .449

Table 3.6.16

Association Between Standardized Residuals and Item Formats (Booklet No. 2, 75 Items, 9 Year Olds, 1977-78)

Format	Standardized Residuals	1-p Results N	1-p Results %	3-p Results N	3-p Results %
Multiple-Choice	SR($\leq$1.0)	8	10.7	32	42.7
	SR(>1.0)	32	42.7	8	10.7
Open-Ended	SR($\leq$1.0)	4	5.3	25	33.3
	SR(>1.0)	31	41.3	10	13.3
		χ^2 = .482		χ^2 = .358	
		d.f. = 1	p = .487	d.f. = 1	p = .551

Table 3.6.17

Association Between Standardized Residuals and Item Difficulties
(Booklet No. 2, 75 Items, 9 Year Olds, 1977-78)

Difficulty Level	Standardized Residuals	1-p Results N	1-p Results %	3-p Results N	3-p Results %
Hard ($p<.5$)	SR($\leq$1.0)	6	8.0	31	41.3
	SR($>$1.0)	34	45.3	9	12.0
Easy ($p\geq.5$)	SR($\leq$1.0)	6	8.0	26	34.7
	SR($>$1.0)	29	38.7	9	12.0
		$\chi^2= 0$		$\chi^2= .003$	
		d.f.= 1	p=1.00	d.f.=1	p=.957

Table 3.6.18

Association Between Item Formats
and Item Difficulties
(Booklet No. 2, 75 Items, 9 Year Olds, 1977-78)

Difficulty Level	Format	N	%
Hard ($p<.5$)	Multiple-Choice	23	30.7
	Open-Ended	17	22.7
Easy ($p\geq.5$)	Multiple-Choice	17	22.7
	Open-Ended	18	24.0
		$\chi^2=.293$	
		d.f.=1	p=.588

Table 3.6.19

Descriptive Statistical Analysis of Standardized Residuals (Booklet No. 2, 75 Items, 9 Year Olds, 1977-78)

Difficulty Level	Format	Number of Items	1-p Results		3-p Results	
			$\overline{X}$	SD	$\overline{X}$	SD
Hard ($p<.5$)	Multiple-Choice	23	2.69	1.47	.81	.25
	Open-Ended	17	1.67	.68	.89	.31
Easy ($p\geq.5$)	Multiple-Choice	17	1.81	1.12	1.04	1.15
	Open-Ended	18	1.72	.64	.91	.20

Table 3.6.20

Relationship Between Item Discrimination Indices and Standardized Residuals

(Booklet No. 2, 75 Items, 9 Year Olds, 1977-78)

Model	Standardized Residuals	Discrimination Indices -.01 to .30	 .31 to .50	 .51 to .70	 .71 to 1.00
		(9)[1]	(18)	(34)	(14)
1-p	0.00 to 1.00	0.0	11.1	29.4	0.0
	1.01 to 2.00	0.0	27.8	70.6	14.3
	over 2.00	100.0	61.1	0.0	85.7
		χ^2= 50.77 Eta= .744	d.f.= 6	p=.000	
3-p	0.00 to 1.00	77.8	83.3	73.5	71.4
	1.01 to 2.00	11.1	16.7	26.5	28.6
	over 2.00	11.1	0.0	0.0	0.0
		χ^2= 8.78 Eta= .114	d.f.=6	p=.186	

[1] Number of test items appear in brackets.

Table 3.6.21

Correlations Among Several NAEP Math Item Variables
(Booklet No. 1, 58 Items, 13 Year Olds, 1977-78)

Variable	SR(1-p)	SR(3-p)	p	r	F[1]
Item Order	.06	-.03	-.27	-.20	-.09
Standardized Residual (1-p)		.16	-.38	-.57	-.09
Standardized Residual (3-p)			-.03	.05	.27
Item Difficulty (p)				.41	.08
Item Discrimination (r)					.46
Format (F)					

[1] 1=Multiple-Choice; 2=Open-Ended.

Table 3.6.22

Association Between Standardized Residuals
and NAEP Item Content Classifications
(Booklet No. 1, 58 Items, 13 Year Olds, 1977-78)

		Standardized Residuals			
		1-p		3-p	
Content Category	Number of Items	SR(≤1.0) (n= 13)	SR(>1.0) (n= 45)	SR(≤1.0) (n= 45)	SR(>1.0) (n= 13)
Story Problems	14	21.4	78.6	85.7	14.3
Geometry	14	28.6	71.4	85.7	14.3
Definitions	9	33.3	66.7	88.9	11.1
Calculations	15	13.3	86.7	53.3	46.7
Measurement	5	20.0	80.0	80.0	20.0
Graphs and Figures	1	0.0	100.0	100.0	0.0
		$\chi^2 = 1.95$		$\chi^2 = 7.10$	
		d.f.= 5	p= .856	d.f.= 5	p= .213

Table 3.6.23

Association Between Standardized Residuals and Item Formats (Booklet No. 1, 58 Items, 13 Year Olds, 1977-78)

Format	Standardized Residuals	1-p Results N	1-p Results %	3-p Results N	3-p Results %
Multiple-Choice	SR(≤1.0)	9	15.5	31	53.4
	SR(>1.0)	27	46.6	5	8.6
Open-Ended	SR(≤1.0)	4	6.9	14	24.1
	SR(>1.0)	18	31.0	8	13.8
		χ^2=.078		χ^2=2.78	
		d.f.=1	p=.780	d.f.=1	p=.096

Table 3.6.24

Association Between Standardized Residuals and Item Difficulties

(Booklet No. 1, 58 Items, 13 Year Olds, 1977-78)

Difficulty Level	Standardized Residuals	1-p Results N	%	3-p Results N	%
Hard ($p<.5$)	SR($\leq$1.0)	2	3.4	19	32.8
	SR($>$1.0)	20	34.5	3	5.2
Easy ($p\geq.5$)	SR($\leq$1.0)	11	19.0	26	44.8
	SR($>$1.0)	25	43.1	10	17.2
		χ^2=2.49		χ^2=.862	
		d.f.=1	p=.115	d.f.=1	p=.353

Table 3.6.25

Association Between Item Formats and Item Difficulties

(Booklet No. 1, 58 Items, 13 Year Olds, 1977-78)

Difficulty Level	Format	N	%
Hard ($p<.5$)	Multiple-Choice	15	25.9
	Open-Ended	7	12.1
Easy ($p\geq.5$)	Multiple-Choice	21	36.2
	Open-Ended	15	25.9
		$\chi^2=.22$	
		d.f.=1	p=.638

Table 3.6.26

Descriptive Statistical Analysis of
Standardized Residuals
(Booklet No. 1, 58 Items, 13 Year Olds, 1977-78)

Difficulty Level	Format	Number of Items	1-p Results		3-p Results	
			$\bar{X}$	SD	$\bar{X}$	SD
Hard ($p<.5$)	Multiple-Choice	15	2.38	1.60	.84	.19
	Open-Ended	7	1.88	1.33	.94	.31
Easy ($p\geq.5$)	Multiple-Choice	21	1.72	1.21	.84	.38
	Open-Ended	15	1.73	.83	1.09	.45

Table 3.6.27

Relationship Between Item Discrimination Indices and Standardized Residuals (Booklet No. 1, 58 Items, 13 Year Olds, 1977-78)

Model	Standardized Residuals	Discrimination Indices -.01 to .30	 .31 to .50	 .51 to .70	 .71 to 1.00
		(6)[1]	(15)	(26)	(11)
1-p	0.00 to 1.00	0.0	13.3	42.3	0.0
	1.01 to 2.00	0.0	53.3	50.0	45.5
	over 2.00	100.0	33.3	7.7	54.5
		$\chi^2 = 26.9$	d.f.= 6	p=.000	
		Eta= .628			
3-p	0.00 to 1.00	66.7	86.7	80.8	63.6
	1.01 to 2.00	33.3	13.3	19.2	36.4
	over 2.00	0.0	0.0	0.0	0.0
		$\chi^2 = 2.51$	d.f.=3	p=.474	
		Eta=.208			

---

[1] Number of test items appear in brackets.

Table 3.6.28

Correlations Among Several NAEP Math Item Variables (Booklet No. 2, 62 Items, 13 Year Olds, 1977-78)

Variable	SR(1-p)	SR(3-p)	p	r	F[1]
Item Order	.21	-.13	-.29	-.16	-.29
Standardized Residual (1-p)		.29	-.43	-.71	-.31
Standardized Residual (3-p)			-.24	-.29	.17
Item Difficulty (p)				.38	.07
Item Discrimination (r)					.44
Format (F)					

[1] 1=Multiple-Choice; 2=Open-Ended.

Table 3.6.29

Association Between Standardized Residuals and NAEP Item Content Classifications (Booklet No. 2, 62 Items, 13 Year Olds, 1977-78)

		Standardized Residuals			
		1-p		3-p	
Content Category	Number of Items	SR(≤1.0) (n= 15)	SR(>1.0) (n= 47)	SR(≤1.0) (n= 47)	SR(>1.0) (n= 15)
Story Problems	15	26.7	73.3	80.0	20.0
Geometry	11	18.2	81.8	63.6	36.4
Definitions	7	0.0	100.0	85.7	14.3
Calculations	22	31.8	68.2	77.3	22.7
Measurement	1	100.0	0.0	0.0	100.0
Graphs and Figures	6	16.7	83.3	83.3	16.7
		$\chi^2 = 6.52$		$\chi^2 = 4.75$	
		d.f.= 5	p= .259	d.f.= 5	p=.447

Table 3.6.30

Association Between Standardized Residuals and Item Formats
(Booklet No. 2, 62 Items, 13 Year Olds, 1977-78)

Format	Standardized Residuals	1-p Results N	%	3-p Results N	%
Multiple-Choice	SR(≤1.0)	5	8.1	26	41.9
	SR(>1.0)	25	40.3	4	6.5
Open-Ended	SR(≤1.0)	10	16.1	21	33.9
	SR(>1.0)	22	35.5	11	17.7
		$\chi^2 = 1.09$		$\chi^2 = 2.67$	
		d.f.=1	p=.297	d.f.=1	p=.102

Table 3.6.31

Association Between Standardized Residuals and Item Difficulties (Booklet No. 2, 62 Items, 13 Year Olds, 1977-78)

Difficulty Level	Standardized Residuals	1-p Results N	1-p Results %	3-p Results N	3-p Results %
Hard ($p<.5$)	SR($\leq$1.0)	2	3.2	17	27.4
	SR($>$1.0)	23	37.1	8	12.9
Easy ($p\geq.5$)	SR($\leq$1.0)	13	21.0	30	48.4
	SR($>$1.0)	24	38.7	7	11.3
		χ^2=4.60		χ^2=.77	
		d.f.=1	p=.032	d.f.=1	p=.380

Table 3.6.32

Association Between Item Formats
and Item Difficulties
(Booklet No. 2, 62 Items, 13 Year Olds, 1977-78)

Difficulty Level	Format	N	%
Hard ($p<.5$)	Multiple-Choice	12	19.4
	Open-Ended	13	21.0
Easy ($p\geq.5$)	Multiple-Choice	18	29.0
	Open-Ended	19	30.6
		$\chi^2 = 0$	
		d.f.= 1	p=1.00

Table 3.6.33

Descriptive Statistical Analysis of
Standardized Residuals
(Booklet No. 2, 62 Items, 13 Year Olds, 1977-78)

Difficulty Level	Format	Number of Items	1-p Results		3-p Results	
			$\bar{X}$	SD	$\bar{X}$	SD
Hard (p<.5)	Multiple-Choice	12	3.07	2.06	.87	.24
	Open-Ended	13	1.59	.72	.86	.31
Easy (p≥.5)	Multiple-Choice	18	1.71	1.04	.74	.16
	Open-Ended	19	1.36	.62	.91	.38

Table 3.6.34

Relationship Between Item Discrimination Indices and Standardized Residuals

(Booklet No. 2, 62 Items, 13 Year Olds, 1977-78)

Model	Standardized Residuals	Discrimination Indices -.01 to .30	 .31 to .50	 .51 to .70	 .71 to 1.00
		(4)[1]	(9)	(36)	(13)
1-p	0.00 to 1.00	0.0	11.1	38.6	0.0
	1.01 to 2.00	0.0	11.1	55.6	46.2
	over 2.00	100.0	77.8	5.6	53.8
		χ^2= 34.40 Eta= .666	d.f.= 6	p= .000	
3-p	0.00 to 1.00	50.0	77.8	77.8	76.9
	1.01 to 2.00	50.0	22.2	22.2	23.1
	over 2.00	0.0	0.0	0.0	0.0
		χ^2= 1.56 Eta= .158	d.f.= 3	p= .669	

---

[1] Number of test items appear in brackets.

Table 3.6.35

Correlations Among Several NAEP Math Item Variables
(Booklet No. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78)

Variable	SR(1-p)	SR(3-p)	p	r	F[1]
Item Order	--	--	--	--	--
Standardized Residual (1-p)		.18	-.30	-.62	-.25
Standardized Residual (3-p)			.09	-.11	.07
Item Difficulty (p)				.41	.04
Item Discrimination (r)					.49
Format (F)					

[1] 1=Multiple-Choice; 2=Open-Ended.

Table 3.6.36

Association Between Standardized Residuals and NAEP Item Content Classifications
(Booklets No. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78)

		Standardized Residuals			
		1-p		3-p	
Content Category	Number of Items	SR(≤1.0) (n= 48)	SR(>1.0) (n=212)	SR(≤1.0) (n=197)	SR(>1.0) (n= 63)
Story Problems	52	21.2	78.8	82.7	17.3
Geometry	48	22.9	77.1	75.0	25.0
Definitions	42	16.7	83.3	78.6	21.4
Calculations	83	15.7	84.3	69.9	30.1
Measurement	17	11.8	88.2	82.4	17.6
Graphs and Figures	18	22.2	77.8	72.2	27.8
		$\chi^2 = 2.08$		$\chi^2 = 3.65$	
		d.f.= 5	p= .838	d.f.= 5	p= .602

Table 3.6.37

Association Between Standardized Residuals and Item Formats (Booklets No. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78)

Format	Standardized Residuals	1-p Results N	1-p Results %	3-p Results N	3-p Results %
Multiple-Choice	SR(≤1.0)	24	9.2	115	44.2
	SR(>1.0)	116	44.6	25	9.6
Open-Ended	SR(≤1.0)	24	9.2	82	31.5
	SR(>1.0)	96	36.9	38	14.6
		$X^2 = .186$		$X^2 = 5.98$	
		d.f.= 1	p= .666	d.f.= 1	p= .015

Table 3.6.38

Association Between Standardized Residuals and Item Difficulties

(Booklets No. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78)

Difficulty Level	Standardized Residuals	1-p Results		3-p Results	
		N	%	N	%
Hard ($p<.5$)	SR($\leq$1.0)	14	5.4	99	38.1
	SR($>$1.0)	110	42.3	25	9.6
Easy ($p\geq.5$)	SR($\leq$1.0)	34	13.1	98	37.7
	SR($>$1.0)	102	39.2	38	14.6
		χ^2 = 7.21		χ^2 = 1.74	
		d.f.= 1	p= .007	d.f.= 1	p= .188

Table 3.6.39

Association Between Item Formats and Item Difficulties (Booklets No. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78)

Difficulty Level	Format	N	%
Hard ($p<.5$)	Multiple-Choice	70	26.9
	Open-Ended	54	20.8
Easy ($p\geq.5$)	Multiple-Choice	70	26.9
	Open-Ended	66	25.4

$\chi^2 = .463$

d.f.= 1 p= .496

Table 3.6.40

Descriptive Statistical Analysis of Standardized Residuals

(Booklets No. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78)

Difficulty Level	Format	Number of Items	1-p Results		3-p Results	
			$\overline{X}$	SD	$\overline{X}$	SD
Hard (p<.5)	Multiple-Choice	70	2.73	1.55	.82	.23
	Open-Ended	54	1.64	.81	.86	.28
Easy (p≥.5)	Multiple-Choice	70	1.79	1.10	.90	.64
	Open-Ended	66	1.67	.72	.97	.38

Table 3.6.41

Relationship Between Item Discrimination Indices and Standardized Residuals

(Booklets No. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78)

Model	Standardized Residuals	Discrimination Indices -.01 to .30	 .31 to .50	 .51 to .70	 .71 to 1.00
		(29)[1]	(55)	(125)	(51)
1-p	0.00 to 1.00	0.0	10.9	33.6	0.0
	1.01 to 2.00	0.0	32.7	62.4	29.4
	over 2.00	100.0	56.4	4.0	70.6
		$\chi^2 = 143.7$ Eta= .691	d.f.= 6	p= 0	
3-p	0.00 to 1.00	75.9	80.0	76.8	68.6
	1.01 to 2.00	20.7	18.2	23.2	29.4
	over 2.00	3.4	1.8	0.0	2.0
		$\chi^2 = 5.28$ Eta= .092	d.f.= 6	p= .508	

---

[1] Number of test items appear in brackets.

Table 3.6.36

- Relationship between standardized residuals and content categories.

1. The pattern of standardized residuals is the same across content categories. Misfit statistics for both the one- and three-parameter models clearly are unrelated to the content of the test items. Of course, the standardized residuals are substantially smaller for the three-parameter model.

Table 3.6.37

- Association between standardized residuals and item formats. It seemed useful to know whether the pattern of misfit statistics for multiple-choice and open-ended test items was the same with the one- and three-parameter models.

1. The pattern of misfit statistics with the one-parameter model is about the same with the two item formats. Residuals were somewhat larger with multiple-choice items.

2. The pattern of misfit statistics with the three-parameter model was also about the same for the two item formats. Somewhat surprisingly the results were a little poorer with the open-ended items. One conjecture is that the c parameters were over estimated.[1]

Table 3.6.38

- Associations between standardized residuals and item difficulty.

1. The one-parameter standardized residuals were substantially higher for difficult items than for easy items.

2. The three-parameter standardized residuals were unrelated to item difficulty.

Table 3.6.39

- Association between item formats and item difficulty.

1. There were approximately the same number of hard and easy test items, and the distribution of items in each format for hard and easy items was about the same. There were a few more easy open-ended test items than hard open-ended test items.

---

[1] The problem was likely due to our failure to designate some test items as "open-ended" when running LOGIST.

Table 3.6.40

- Analysis of standardized residuals for items organized by difficulty and format.

1. Hard multiple-choice items had substantially larger residuals when fit by the one-parameter model than easy items in either format, or hard items in open-ended format. This result suggests that the problem is due to a failure to account for guessing behavior (note, the fit was better for hard open-ended items where guessing behavior is not operative). The differences between the average one-parameter and three-parameter model standardized residuals, except for the hard multiple-choice test items, are probably due to the difference in the way item discriminating power is handled. With the hard multiple-choice test items, the difference is due to a failure to account for both item discriminating power and examinee guessing behavior in the one-parameter model.

2. There were no relationships among item difficulty level, item format, and standardized residuals obtained from fitting the three-parameter model.

Table 3.6.41

- Relationship between item discrimination indices and standardized residuals.

1. The one-parameter model residuals are non-linearly related to classical item discrimination indices (Eta=.691).

2. The three-parameter model residuals are not related in any fashion to classical item discrimination indices (Eta=.061).

In summary, the results of our hypothesis testing showed clearly that the test items in the content categories we worked with were not in any way being fit better or worse by the item reponse models, and failure to consider examinee guessing behavior and variation in item discriminating power resulted in the one-parameter model providing substantially poorer fits to the various test data sets than the three-parameter model.

4.0 Conclusions

4.1 Implications of Findings for NAEP

The potential of item response theory has been widely documented but that potential is certainly not guaranteed when applied to particular tests, with particular samples of examinees, or when used in particular applications. Item response theory is not a magic wand to wave over a data set to fix all of the inaccuracies and inadequacies in a test and/or the testing procedures. But, when a bank of content valid and technically sound test items is available, and goodness of fit studies reveal a high match between the chosen item response model and the test data, item response models may be useful to NAEP in test development, detection of biased items, score reporting, equating test forms and levels, item banking, and other applications as well. The goals of this study were in a general way aimed at all possible item response model applications to NAEP data, but specifically aimed at the possible uses of item response models in mathematics item banking, one of the lesser important concerns of ECS on the NAEP project at the present time. Still, there is great interest at the national, state, district, and school-level in item banking and NAEP exercises. In addition to the overall quality of NAEP exercises, NAEP exercises are "normed" and so interest in them and their statistics is high.

The implications of the present study for NAEP are the following:

1. A large number of goodness of fit investigations were described in Chapter 2 and several new investigations were conducted and described in Chapter 3. Many of these investigations can now be tried on other NAEP

data sets to determine the generalizability of the conclusions drawn in this study concerning model data fit.

2. The findings of this investigation clearly support the desirability of conducting a wide range of analyses on a data set, and on several data sets. Were a narrow set of analyses to be conducted on (possibly) a single model and data set the interpretation of results would have been more confusing and difficult. The approaches described in Figure 2.3.1 should provide some direction to NAEP staff and other researchers with an interest in IRT applications.

3. It seems clear that the three-parameter model performed substantially better than the one-parameter model. The results were not especially surprising, given information about the ways in which the NAEP exercises are constructed (i.e., relatively little use is made of item statistical information in test development). While the utility of the three- over the one-parameter model was not too surprising, the actual fits of the three-parameter model to the data sets were. The study of standardized residuals at the item level and ability level revealed a very good fit of the three-parameter model.

4. Not all of the analyses revealed high three-parameter model-test data fit. The studies of "bias" were the most confusing. Regardless of whether the three-parameter model or the one-parameter model was fitted to the data, a number of potentially "biased" items were identified. Several possible explanations exist: Several test items are biased against one group or another (e.g., race, or high and low performers) or there are problems in item parameter estimation (e.g., c parameters cannot be properly estimated in high performing groups, or in any groups — black or white or hispanic — if group size is of the size used in this investigation).

5. Perhaps the most important finding is that it is highly unlikely that the one-parameter model will be useful with NAEP mathematics exercises. This is in spite of the fact that many other organizations are very pleased with their work with the one-parameter model. With NAEP mathematics booklets it appears there is too much variation among mathematics items in their discriminating power and too much guessing on the hard multiple-choice test items for the one-parameter model to provide an adequate fit to the test data.

It is our opinion that the results from the first part of the study will be of interest and value to measurement specialists who are considering the usefulness of item response models in their work. Essentially, we are recommending that measurement specialists design and carry out a comprehensive set of analyses to provide themselves with sufficient information to make informed judgments about the usefulness of item response models in their particular applications. The amount of effort extended in collecting information will be, of course, directly related to the importance of the intended applications.

The second part of the study provides information that can impact on the future use of item response models in NAEP. There is considerable evidence in Chapter 3 suggesting that the three-parameter logistic model provides a very good accounting of the actual mathematics test results. The one-parameter logistic model did not. It may be that NAEP will now want to consider utilizing the three-parameter model in some small scale item bias, item banking, and test development efforts to determine the utility and appropriateness of the three-parameter model. Such investigations seem highly worthwhile at this time. Of course, it may be that with other content areas the one-parameter model may suffice, and for problems of score reporting new models being developed by Bock, Mislevy, and Woodson may be substantially better than the three-parameter logistic model.

5.0 References

Andersen, E.B. A goodness of fit test for the Rasch model. *Psychometrika*, 1973, *38*, 123-140.

Baker, F.B. An intersection of test score interpretation and item analysis. *Journal of Educational Measurement*, 1964, *1*, 23-28.

Baker, F.B. Origins of the item parameters X_{50} and β as a modern item analysis technique. *Journal of Educational Measurement*, 1965, *2*, 167-180.

Bejar, I.I. A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement*, 1980, *17*, 283-296.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Bock, R.D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 1972, *37*, 29-51.

Bock, R.D., & Lieberman, M. Fitting a response model for n dichotomously scored items. *Psychometrika*, 1970, *35*, 179-197.

Cronbach, L.J., & Warrington, W.G. Time-limit tests: Estimating their reliability and degree of speeding. *Psychometrika*, 1951, *16*, 167-188.

Divgi, D.R. Does the Rasch model really work? Not if you look closely. Paper presented at the annual meeting of NCME, Los Angeles, 1981.

Donlon, T.F. An exploratory study of the implications of test speededness. Princeton, NJ: Educational Testing Service, 1978.

Green, S.B., Lissitz, R.W., & Mulaik, S.A. Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, 1977, *37*, 827-838.

Gulliksen, H. *Theory of mental tests*. New York: Wiley, 1950.

Hambleton, R.K. Latent trait models and their applications. In R. Traub (Ed.), *Methodological developments: New directions for testing and measurement (No. 4)*. San Francisco: Jossey-Bass, 1980.

Hambleton, R.K. Latent ability scales, interpretations, and uses. In S. Mayo (Ed.), *New directions for testing and measurement: Interpreting test scores (No. 6)*. San Francisco: Jossey-Bass, 1980.

Hambleton, R.K. (Ed.) Applications of item response models. Vancouver, BC: Educational Research Institute of British Columbia, 1982. (a)

Hambleton, R.K. Applications of item response models to criterion-referenced assessments. Applied Psychological Measurement, 1982, 6, in press. (b)

Hambleton, R.K. Advances in criterion-referenced testing technology. In C. Reynolds & T. Gutkin (Eds.), Handbook of School Psychology. New York: Wiley, 1982. (c)

Hambleton, R.K., & Cook, L.L. The robustness of latent trait models and effects of test length and sample size on the precision of ability estimates. In D. Weiss (Ed.), New Horizons in Testing. New York: Academic Press, 1982.

Hambleton, R.K., & Swaminathan, H. Introduction to item response models and their applications. Boston: Martinus-Nijhoff Publishers, 1982.

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R., & Gifford, J.A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Hambleton, R.K., & Traub, R.E. Analysis of empirical data using two logistic latent trait models. British Journal of Mathematical and Statistical Psychology, 1973, 26, 195-211.

Hattie, J.A. Decision criteria for determining unidimensionality. Unpublished doctoral dissertation, University of Toronto, 1981.

Horn, J.L. A rationale and test for the number of factors in factor analysis. Psychometrika, 1965, 30, 179-185.

Linn, R.L., & Harnisch, D.L. Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 1980, 17, 179-194.

Lord, F.M. A theory of test scores. Psychometric Monograph, 1952, No. 7.

Lord, F.M. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. Psychometrika, 1953, 18, 57-76.

Lord, F.M. Estimating item characteristic curves without knowledge of their mathematical form. Psychometrika, 1970, 35, 43-50.

Lord, F.M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

Lord, F.M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

Lord, F.M., & Novick, M.R. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Lumsden, J. The construction of unidimensional tests. *Psychological Bulletin*, 1961, *58*, 122-131.

Lumsden, J. Test theory. *Annual Review of Psychology*, 1976, *27*, 251-280.

McDonald, R.P. The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 1980, *33*, 205-233. (a)

McDonald, R.P. Fitting latent trait models. In D. Spearitt (Ed.), *The Improvement of Measurement in Education and Psychology*. Proceedings of the Invitational Seminar for the Fiftieth Anniversary of the Australian Council of Educational Research, Melbourne, 1980. (b)

McDonald, R.P., & Ahlawat, K.S. Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*, 1974, *27*, 82-99.

Mead, R. Assessing the fit of data to the Rasch model. A paper presented at the annual meeting of AERA, San Francisco, 1976.

Popham, W.J. *Modern educational measurement*. Englewood Cliffs, NJ: Prentice-Hall, 1980.

Reckase, M.D. Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 1979, *4*, 207-230.

Ree, M.J. Estimating item characteristic curves. *Applied Psychological Measurement*, 1979, *3*, 371-385.

Ross, J. An empirical study of a logistic mental test model. *Psychometrika*, 1966, *31*, 325-340.

Samejima, F. A use of the information function in tailored testing. *Applied Psychological Measurement*, 1977, *1*, 233-247.

Swaminathan, H. Bayesian estimation in the two-parameter logistic model. *Laboratory of Psychometric and Evaluative Research Report No. 112*. Amherst, MA: University of Massachusetts, 1981.

van den Wollenberg, A.L. On the Wright-Panchapakesan goodness of fit test for the Rasch model. Nijmegen, The Netherlands: Department of Mathematical Psychology, University of Nijmegen, 1980.

van den Wollenberg, A.L. A simple and effective method to test the dimensionality axiom of the Rasch model. *Applied Psychological Measurement*, 1982, in press. (a) van den Wollenberg, A.L. Two new test statistics for the Rasch model. Psychometrika, 1982, in press. (b)

Waller, M.I. A procedure for comparing logistic latent trait models. Journal of Educational Measurement, 1981, 18, 119-125.

Wingersky, M.S. LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R.K. Hambleton (Ed.), Applications of Item Response Models. Vancouver, BC: Educational Research Institute of British Columbia, 1982.

Wingersky, M.S., Barton, M.A., & Lord, F.M. LOGIST user's guide. Princeton, NJ: Educational Testing Service, 1982.

Wright, B.D. Sample free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service, 1968.

Wright, B.D., Mead, R., & Draba, R. Detecting and correcting item bias with a logistic response model. Research Memorandum No. 22. Chicago: University of Chicago, Statistical Laboratory, Department of Education, 1976.

Wright, B.D., & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-37.

Wright, B.D., & Stone, M.H. Best test design. Chicago: MESA, 1979.

Yen, W.M. The extent, causes and importance of context effects on item parameters for two latent trait models. Journal of Educational Measurement, 1980, 17, 297-311.

Appendix A

Item Response Model Goodness of Fit Studies

Item Response Model Goodness of Fit Studies[1]

Andersen, E.B. A goodness of fit test for the Rasch model. Psychometrika, 1973, 38, 123-140.

Bejar, I.I. A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 1980, 17, 283-296.

Baker, F.B. The effect of criterion score grouping upon item parameter estimation. British Journal of Mathematical and Statistical Psychology, 1967, 20, 227-238.

Bentler, P.M. and Bonett, D.G. Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 1980, 88, 588-606.

Callender, J.C. and Osburn, H.G. An empirical comparison of coefficient Alpha, Guttman's lambda-2, and msplit maximized split-half reliability estimates. Journal of Educational Measurement, 1979, 16, 89-99.

Cattell, R.B. The scree test for the number of factors. Multivariate Behavioral Research, 1966, 1, 245-276.

Cattell, R.B. and Vagelmann, S. A comprehensive trial of the score and kg criteria for determining the number of factors. Multivariate Behavioral Research, 1977, 12, 289-325.

Christoffersson, A. Factor analysis of dichotomized variables. Psychometrika, 1975, 40, 5-32.

Crane, J.A. Relative likelihood analysis versus significance tests. Evaluation Review, 1980, 4, 824-842.

Donlon, T.F. An exploratory study of the implications of test speededness. Unpublished manuscript, 1978.

Frisbie, D.A. A method for comparing test difficulties. A paper presented at the annual meeting of the National Council on Measurement and Education, Los Angeles, 1981.

George, A.A. Theoretical and practical consequences of the use of standardized residuals as Rasch model fit statistics. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Goldstein, H. Changing educational standards: a fruitless search. Journal of the National Association of Inspectors and Educational Advisors, 1979, 11, 18-19.

---

[1] Prepared by Ronald K. Hambleton and Linda Murray.

Goldstein, H. Dimensionality, bias, independence and measurement scale problems in latent trait test score models. British Journal of Mathematical and Statistical Psychology, 1980, 33, 234-246.

Goldstein, H. Consequences of using the Rasch model for educational assessment. British Educational Research Journal, 1979, 5, 211-220.

Green, S.B., Lissitz, R.W., and Mulaik, S.A. Limitations of coefficient alpha as an index of test unidimensionality. Educational and Psychological Measurement, 1977, 37, 827-838.

Gustafsson, J.E. Testing and obtaining fit of data to the Rasch model. British Journal of Mathematical and Statistical Psychology, 1980, 33, 205-233.

Hakstian, A.R. and Muller, V.J. Some notes on the number of factors problem. Multivariate Behavioral Research, 1973, 8, 461-475.

Hambleton, R.K. and Traub, R.E. Analysis of empirical data using two logistic latent trait models. British Journal of Mathematical and Statistical Psychology, 1973, 26, 195-211.

Hambleton, R.K. and Cook, L.L. Some results on the robustness of latent trait models. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, 1978.

Hartwig, F. and Dearing, B.E. Explatory Data Analysis. Beverly Hills, CA: Sage Publications, 1979.

Holland, P.W. When are item response models consistent with observed data? Psychometrika, 1981, 46, 79-92.

Horn, J.L. A rationale and test for the number of factors in factor analysis. Psychometrika, 1965, 30, 179-185.

Humphreys, L.G. and Montanelli, R.G. An investigation of the parallel analysis criterion for determining the number of common factors. Multivariate Behavioral Research, 1975, 10, 193-205.

Jöreskog, K.G. Estimation and testing of simplex models. British Journal of Mathematical and Statistical Psychology, 1970, 23, 121-145.

Linn, R.L. and Harnisch, D.L. Interaction between item content and group membership on achievement test items. Journal of Educational Measurement, 1981, 18, 109-118.

Lord, F.M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F.M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-263.

Lord, F.M. A broad-range tailored test of verbal ability. Applied Psychological Measurement, 1977, 1, 95-100.

Lord, F.M. Item characteristic curves estimated without knowledge of their mathematical form - a confrontation of Birnbaum's logistic model. *Psychometrika*, 1970, *35*, 43-50.

McDonald, R.P. Some alternative approaches to the improvement of measurement in education and psychology: fitting latent trait models. Paper presented at the A.C.E.R. Invitational Seminar, Melbourne, May 22-23, 1980.

McDonald, R.P. The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 1981, *34*, 100-117.

Miller, J. and Greeno, J.G. Goodness-of-fit tests for models of latency and choice. *Journal of Mathematical Psychology*, 1978, *17*, 1-13.

Mukherjee, B.N. Derivation of likelihood-ratio tests for Guttman quasi-simplex covariance structures. *Psychometrika*, 1966, *31*, 97-124.

Reiser, M.R. A latent trait model for group effects. Unpublished dissertation, University of Chicago, 1980.

Slinde, J.A. and Linn, R.L. The Rasch model, objective measurement, equating, and robustness. *Applied Psychological Measurement*, 1979, *3*, 437-452.

Terwilliger, J.S. and Lele, K. Some relationships among internal consistency, reproducibility, and homogeneity. *Journal of Educational Measurement*, 1979, *16*, 101-108.

Traub, R.E. and Wolfe, R.G. Latent trait theories and the assessment of educational achievement. In D. Berliner (Ed.), *Review of Research in Education - Vol. 9*. Washington: American Educational Research Association, 1981.

Tucker, L.R. and Lewis, C. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 1973, *38*, 1-10.

Waller, M.I. A procedure for comparing logistic latent trait models. *Journal of Educational Measurement*, 1981, *18*, 119-125.

Whitely, S.E. Models, meanings, and misunderstandings: some issues in applying Rasch's theory. *Journal of Educational Measurement*, 1977, *14*, 227-235.

Wollenberg, A.L. van den. The Rasch model and time-limit tests: an application and some theoretical contributions. Unpublished dissertation, University of Nijmegen, 1979.

Wright, B.D. Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service, 1968.

Wright, B.D. Misunderstanding the Rasch model. *Journal of Educational Measurement*, 1977, *14*, 219-225.

Appendix B

Item Response Model Residual Analysis Program
(Program Listing and Sample Output)

```
C
C THIS PROGRAM WAS DESIGNED AT THE UNIVERSITY OF MASSACHUSETTS
C      ON THE CDC CYBER 175 IN THE FORTRAN VERSION 5 LANGUAGE
C
C
C
C
      PROGRAM TEST(INPUT,OUTPUT,TAPE50,TAPE60,TAPE8,TAPE5)
      DIMENSION  EX(19,75),A(75),B(75),C(75)
      DIMENSION  SE(19,75),OE(19,75),SV(19,75),RESID(19,75)
      DIMENSION  IRESULT(19,76),IANS(75),PRESULT(19,75)
      DIMENSION  ITITLE(20),ABIL(19)
      REAL LEVEL
      CHARACTER*10 DATE,TODAY
C
C     OUTPUT IS PRINTED ON TAPE 8
C DATA  IS READ IN ON TAPE5, TAPE50, TAPE60
C TAPE50 CONTAINS ITEM PARAMETERS
C TAPE60 CONTAINS ABILITY ESTIMATES AND RESPONSE VECTORS FOR ALL EXAMINES
C TAPE5 CONTAINS PROGRAM OPTIONS
C
C
C DIRECTIONS FOR SETTING UP DATA DECK ON TAPE5
C
C
C CARD 1-READ IN USER DEFINED TITLE (ITITLE)
C CARD 2-NUMBER OF ITEMS (ITEMS)
C CARD 3-NUMBER OF EXAMINEES (NSUBJ)
C CARD 4-MAXIMUM ABILITY VALUE (RMAX)
C CARD 5-MINIMUM ABILITY VALUE (RMIN)
C CARD 6-SIZE OF AN ABILITY CATEGORY (SIZINT)
C CARD 7-DO YOU WANT THI P VALUES (IPP) PRINTED ?
C        Y FOR YES OR N FOR NO
C CARD 8-DO YOU WANT THE RESIDUALS (IPR) PRINTED ?
C        Y FOR YES OR N FOR NO
C CARD 9-DO YOU WANT THE STANDARDIZED RESIDUALS (IPS) PRINTED?
C        Y FOR YES OR N FOR NO
C
      READ (5,520)ITITLE
  520 FORMAT(20A4)
C
C     WRITE HEADING PAGE FOR UMASS
C
      TODAY=DATE()
      WRITE(8,11)TODAY
   11 FORMAT(1H1,///////,35X,"RESIDUAL ANALYSES OF LOGISTIC TEST DATA",
     */,50X,"DATE:",A10,/,50X,"(VERSION 3A)",////,43X,
     *"PROGRAM BY LINDA MURRAY",/,51X,"RONALD HAMBLETON",
     */,51X,"ROBERT SIMON",/////,35X,
     *"DEVELOPED AT THE UNIVERSITY OF MASSACHUSETTS",/,44X,
     *"SCHOOL OF EDUCATION",/,42X,
     *"UNDER A GRANT FROM NAEP")
      WRITE(8,530)ITITLE
  530 FORMAT(////,35X,20A4,///)
      READ(5,1)ITEMS
      READ(5,1)NSUBJ
    1 FORMAT(I7)
      READ(5,2)RMAX
    2 FORMAT(F5.2)
      READ(5,2)RMIN
      READ(5,2)SIZINT
      READ(5,3)IPP
      READ(5,3)IPR
    3 FORMAT (A1)
      READ(5,3)IPS
C
```

```
C LIMIT IS THE NUMBER OF ABILITY CATEGORIES
C
      LIMIT=(ABS(RMIN)+RMAX)/SIZINT
C
      WRITE(8,21)ITEMS,NSUBJ,LIMIT
   21 FORMAT(/////,35X,I3," ITEMS / ",I5, " EXAMINEES / ",
     *I2," ABILITY GROUPINGS ")
C
C MODEL 1 IS THE ONE PARAMETER LOGISTIC MODEL
C
C
      MODEL=1
C
C ITEMONE IS THE LAST COLUMN IN THE OBSERVED P VALUE MATRIX
C     FOR COUNTING THE NUMBER OF PEOPLE IN AN ABILITY CATEGORY
C
      ITEMONE=ITEMS+1
C
C READ IN FOR MODEL 1 ON TAPE60 A,B,C PARAMETERS FOR ALL ITEMS
C
      DO 4 I=1,ITEMS
      READ(60,200) A(I),B(I),C(I)
  200 FORMAT (5X,3(F6.3))
    4 CONTINUE
      GO TO 10
    5 CONTINUE
      REWIND 60
C
C READ IN FOR MODEL 3 ON TAPE60 THE A,B,C PARAMETERS FOR ALL ITEMS
C
C MODEL 3 IS THE THREE PARAMETER LOGISTIC MODEL
C
      MODEL=3
      DO 6 I=1,ITEMS
      READ(60,7) A(I),B(I),C(I)
    7 FORMAT(24X,3(F6.3))
    6 CONTINUE
      GO TO 10
    8 CONTINUE
      STOP
C
C THIS IS THE TERMINATION OF THE PROGRAM
C THE CODE BELOW THIS POINT IS USED TO CALCULATE THE P VALUES AND RESIDUALS
C
C
   10 CONTINUE
C
C AB IS THE MIDPOINT OF THE LOWEST ABILITY CATEGORY
C
      AB=RMIN+(SIZINT/2.0)
C
C THIS LOOP CALCULATES THE MIDPOINT OF THE ABILITY CATEGORIES, ABIL(J),
C      AND THE EXPECTED P VALUES MATRIX, EX(J,I)
C
      DO 300 I=1,ITEMS
      DO 300  J=1,LIMIT
      ABIL(J)=AB+(J-1)*SIZINT
      DD=(1.7*A(I)*(ABIL(J)-B(I)))
      D=2.7182818**(DD)
      EX(J,I)=C(I)+(1.0-C(I))*(D/(1.0+D))
      IF(EX(J,I).LT..01)EX(J,I)=.01
      IF(EX(J,I).GT..99)EX(J,I)=.99
  300 CONTINUE
C
```

```
C SET EQUAL TO ZERO THE COUNTER FOR DETERMINING THE NUMBER OF EXAMINEES
C       WITH ABILITY ESTIMATES BEYOND THE CHOSEN RANGE
C
      KCOUNT=0
C
C ZERO OUT ARRAY IRESULT WHICH CONTAINS THE NUMBER OF EXAMINEES IN AN ABILITY CATEGORY
C    AND THE NUMBER OF EXAMINEES WHO GET THE ITEMS CORRECT
C
      DO 480 I=1,LIMIT
      DO 480 J=1,ITEMONE
      IRESULT(I,J)=0
  480 CONTINUE
C
C READ IN FOR EITHER MODEL ON TAPE50 THE ABILITY ESTIMATE (LEVEL) AND THE RESPONSE
C       VECTOR FOR EACH EXAMINEE (IANS(K))
C
      DO 100 I=1,NSUBJ
      IF(MODEL.EQ.3)GO TO 15
      READ(50,490)LEVEL,(IANS(K),K=1,ITEMS)
      GO TO 19
   15 CONTINUE
      READ(50,16)LEVEL,(IANS(K),K=1,ITEMS)
   16 FORMAT(/,10X,F10.3,2X,73I1)
   19 CONTINUE
  490 FORMAT(/,F10.3,10X,2X,73I1)
C
C KCOUNT COUNTS UP THE NUMBER OF EXAMINEES THAT FALL BEYOND THE MAXIMUM
C       OR MINIMUM ABILITY VALUES
C
      IF((LEVEL.GT.RMAX).OR.(LEVEL.LT.RMIN))KCOUNT=KCOUNT+1
      IF((LEVEL.GT.RMAX).OR.(LEVEL.LT.RMIN))GO TO 100
C
C IABIL IS THE ABILITY ESTIMATES TRANSFORMED INTO AN ABILTY CATEGORY
C
      IABIL=(((LEVEL+ABS(RMIN))/SIZINT)+1)
      IRESULT(IABIL,ITEMONE)=IRESULT(IABIL,ITEMONE)+1
      DO 20 J=1,ITEMS
      IRESULT(IABIL,J)=IRESULT(IABIL,J)+IANS(J)
   20 CONTINUE
  100 CONTINUE
C
C PRESULT IS THE MATRIX OF OBSERVED P VALUES
C
      DO 220 I=1,ITEMS
      DO 220 J=1,LIMIT
      PRESULT(J,I)=(FLOAT(IRESULT(J,I))/FLOAT(IRESULT(J,ITEMONE)))
  220 CONTINUE
      REWIND 50
C
C DO YOU WANT TO PRINT OUT THE P VALUE TABLES????????????
C IF SO IPP IS = TO Y
C
      IF(IPP.NE."Y")GO TO 471
C
C PRINT THE EXPECTED P VALUES MATRIX
C
      WRITE(8,250)MODEL
      WRITE(8,110)
  110 FORMAT(55X,'ABILITY LEVEL')
      WRITE(8,120)(I,I=1,LIMIT)
  120 FORMAT(3X,"CATEGORY",1X,19(3X,I3)/)
  250 FORMAT(1H1,////,40X,'EXPECTED P VALUES-',I1,' PARAMETER MODEL',////
     *)
      WRITE(8,55) (ABIL(J),J=1,LIMIT)
   55 FORMAT(/,3X,"MID-POINT",2X,19(1X,F5.2)/)
```

```
      WRITE(8,130) (IRESULT(I,ITEMONE),I=1,LIMIT)
      WRITE(8,150)
  150 FORMAT(/,5X,'ITEM')
      L=10
      DO 400 I=1,ITEMS
      WRITE(8,450)I,(EX(J,I),J=1,LIMIT)
  450 FORMAT(6X,I2,6X,19(F6.3))
      K=MOD(I,L)
      IF(K.NE.0)GO TO 451
      WRITE(8,262)
  451 CONTINUE
  400 CONTINUE
C
C
C PRINT THE OBSERVED P VALUES MATRIX
C
      WRITE(8,470)
  470 FORMAT(////)
      WRITE(8,12)
      WRITE(8,510)MODEL
  102 WRITE(8,110)
  115 WRITE(8,120)(I,I=1,LIMIT)
      WRITE(8,55) (ABIL(J),J=1,LIMIT)
      WRITE(8,130)(IRESULT(I,ITEMONE),I=1,LIMIT)
  130 FORMAT(3X,'NO. OF',/,2X,' EXAMINEES',1X,19(3X,I3)//)
      WRITE(8,150)
  170 FORMAT(////,'THE NO. OF EXCLUDED CASES=',I9)
   12 FORMAT(1H1)
  510 FORMAT(////,40X,'OBSERVED P VALUES-',I1,' PARAMETER MODEL',////)
      DO 270 I=1,ITEMS
      WRITE(8,260)I,(PRESULT(J,I),J=1,LIMIT)
  260 FORMAT(6X,I2,6X,19(F6.3))
      K=MOD(I,L)
      IF(K.NE.0)GO TO 261
      WRITE(8,262)
  262 FORMAT(/)
  261 CONTINUE
  270 CONTINUE
      WRITE(8,170)KCOUNT
C
  471 CONTINUE
C GO DO RESIDUALS AND STD RESIDUALS FOR THIS CASE
C
      CALL R(ITEMS,LIMIT,EX,IRESULT,PRESULT,ITEMONE,MODEL,
     *IFR,IPS,ABIL,NSUBJ,KCOUNT)
C
C IF ONE PARAMETER MODEL GO BACK AND DO THREE PARAMETER MODEL
      IF (MODEL.EQ.1) GOTO 5
C IF THREE PARAMETER MODEL GO TO END OF PROGRAM
      IF (MODEL.EQ.3) GOTO 8
      END
C
C THIS SUBROUTINE CALCULATES RESIDUALS, STD RESIDUALS AND VARIOUS STATISTICS
C
      SUBROUTINE R(ITEMS,LIMIT,EX,IR,PR,ITEMONE,MODEL,
     *IFS,IFR,ABIL,NSUBJ,KCOUNT)
      DIMENSION SE(19,75),OE(19,75),SV(19,75),RESID(19,75)
      DIMENSION IR(19,75),PR(19,75),EX(19,75),ABIL(19)
      IRE0=0
      IRE1=0
      IRE2=0
      IRE3=0
      L=10
C
```

```
C SV IS THE VARIANCE OF THE EXPECTED P VALUES
C SE IS THE STANDARD ERROR OF THE EXPECTED P VALUES
C OE IS THE RESIDUAL=OBSERVED-EXPECTED
C RESID IS THE RESIDUAL STANDARDIZED
C
      DO 900 I=1,ITEMS
      DO 900 J=1,LIMIT
      SV(J,I)=(EX(J,I)*(1.0-EX(J,I)))/FLOAT(IR(J,ITEMONE))
      SE(J,I)=SQRT(SV(J,I))
      OE(J,I)=(PR(J,I)-(EX(J,I)))
      RESID(J,I)=OE(J,I)/SE(J,I)
      IF((ABS(RESID(J,I)).GE.0.000).AND.(ABS(RESID(J,I)).LT.1.000))
     *IRE0=IRE0+1
      IF ((ABS(RESID(J,I)).GE.1.000).AND.(ABS(RESID(J,I)).LT.2.000))
     *IRE1=IRE1+1
      IF ((ABS(RESID(J,I)).GE.2.000).AND.(ABS(RESID(J,I)).LT.3.000))
     *IRE2=IRE2+1
      IF(ABS(RESID(J,I)).GE.3.000)IRE3=IRE3+1
  900 CONTINUE
      PER0=(FLOAT(IRE0)/FLOAT(ITEMS*LIMIT))*100.00
      PER1=(FLOAT(IRE1)/FLOAT(ITEMS*LIMIT))*100.00
      PER2=(FLOAT(IRE2)/FLOAT(ITEMS*LIMIT))*100.00
      PER3=(FLOAT(IRE3)/FLOAT(ITEMS*LIMIT))*100.00
      IF(IPR.NE."Y")GOTO 915
C
C     PRINT RESIDUAL MATRIX
C
C    PRINT RESIDUAL HEADING
C
      WRITE(8,901)MODEL
  901 FORMAT(1H1,/,40X,'RESIDUALS- ',I1,' PARAMETER MODEL',//,
     *45X,'(OBSERVED-EXPECTED)',///)
      WRITE(8,110)
      WRITE(8,911)(I,I=1,LIMIT)
      WRITE(8,1735)(ABIL(J),J=1,LIMIT)
 1735 FORMAT(/,3X,'MID-POINT',2X,19(F7.3)/)
      WRITE(8,1736)(IR(I,ITEMONE),I=1,LIMIT)
 1736 FORMAT(3X,'NO. OF',/,2X,' EXAMINEES',2X,19(2X,I3,2X)//)
      WRITE(8,150)
C
C    PRINT RESIDUAL TABLE OF VALUES
C
      DO 910 I=1,ITEMS
      WRITE(8,260)I,(OE(J,I),J=1,LIMIT)
      K=MOD(I,L)
      IF(K.NE.0)GO TO 909
      WRITE(8,914)
  909 CONTINUE
  910 CONTINUE
C
C    EVALUATE STATISTICS ON THE RESIDUALS
C
      CALL STATS(OE,LIMIT,ITEMS,ABIL,IR,ITEMONE,NSUBJ,KCOUNT)
      WRITE(8,470)
  915 IF(IPS.NE."Y") GOTO 940
C
C    PRINT STANDARDIZED RESIDUAL MATRIX
C
C   PRINT HEADING FOR STD RESIDUALS
C
      WRITE(8,912) MODEL
  912 FORMAT(1H1,////,40X,'STANDARDIZED RESIDUALS- ',I1,'
     *PARAMETER MODEL',////)
      WRITE(8,110)
      WRITE(8,911)(I,I=1,LIMIT)
      WRITE(8,1735) (ABIL(J),J=1,LIMIT)
      WRITE(8,1736) (IR(I,ITEMONE),I=1,LIMIT)
      WRITE(8,150)
  911 FORMAT(14X,19(2X,I3,2X)/)
      DO 920 I=1,ITEMS
      WRITE(8,260)I,(RESID(J,I),J=1,LIMIT)
      K=MOD(I,[illegible])
      IF(K.NE.0)GO TO 913
      WRITE(8,914)
  913 CONTINUE
  914 FORMAT(/)
  920 CONTINUE
      WRITE(8,921)
  921 FORMAT(////,10X,'ANALYSIS OF STANDARDIZED RESIDUALS',/,
     *18X,'(ABSOLUTE VALUES)',/,4X,
     *'INTERVAL  NUMBER  PERCENT     CUMULATIVE',/,4X,30X,'PERCENT',/)
      TOTPER=PER0
      WRITE(8,922)IRE0,PER0,TOTPER
  922 FORMAT(6X,'0 TO 1',3X,I4,3X,F6.2,6X,F6.2,/)
      TOTPER1=PER0+PER1
      WRITE(8,923)IRE1,PER1,TOTPER1
  923 FORMAT(6X,'1 TO 2',3X,I4,3X,F6.2,6X,F6.2,/)
      TOTPER2=PER0+PER1+PER2
      WRITE(8,924)IRE2,PER2,TOTPER2
  924 FORMAT(6X,'2 TO 3',3X,I4,3X,F6.2,6X,F6.2,/)
      TOTPER3=PER0+PER1+PER2+PER3
      WRITE(8,925)IRE3,PER3,TOTPER3
  925 FORMAT(6X,'BEYOND 3  ',I4,3X,F6.2,6X,F6.2,/)
C
C   EVALUATE STATISTICS FOR STD RESIDUALS
C
      CALL STATS(RESID,LIMIT,ITEMS,ABIL,IR,ITEMONE,NSUBJ,KCOUNT)
  940 CONTINUE
  11( FORMAT(56X,'ABILITY LEVEL')
  150 FORMAT(/,5X,'ITEM')
  470 FORMAT(////)
  260 FORMAT (5X,I2,5X,19(F7.3))
      END
C
C
C
C
C
      SUBROUTINE STATS(RESID,LIMIT,ITEMS,ABIL,IR,ITEMONE,NSUBJ,KCOUNT)
      DIMENSION AVELIM(19),ABSLIM(19),RMSLIM(19),RESID(19,75)
      DIMENSION AVEITM(75),ABSITM(75),RMSITM(75),ABIL(19)
      DIMENSION IR(19,75),WAVITM(75),WABITM(75)
      WRITE(8,471)
C
C ZERO THE VALUES FOR STATISTICS
C
       DO 1010 I=1,75
       AVELIM(I)=0
       ABSLIM(I)=0
       RMSLIM(I)=0
       AVEITM(I)=0
       ABSITM(I)=0
       RMSITM(I)=0
       WAVITM(I)=0
       WABITM(I)=0
 1010 CONTINUE
      AVEAVL=0
      AVEABL=0
      AVEAVI=0
      AVEABI=0
      AVERML=0
      AVERMI=0
      AVEWAV=0
      AVEWAB=0
C
C CALCULATE FOR AVERAGE ABILITY LEVEL AVERAGE ABSOLUTE ABILITY LEVEL
C
      DO 1050 I=1,LIMIT
      DO 1000 J=1,ITEMS
     AVELIM(I)=AVELIM(I)+RESID(I,J)
     ABSLIM(I)=ABSLIM(I)+ABS(RESID(I,J))
 1000 CONTINUE
      AVELIM(I)=AVELIM(I)/FLOAT(ITEMS)
      ABSLIM(I)=ABSLIM(I)/FLOAT(ITEMS)
      AVEAVL=AVEAVL+AVELIM(I)
      AVEABL=AVEABL+ABSLIM(I)
 1050 CONTINUE
      AVEAVL=AVEAVL/FLOAT(LIMIT)
      AVEABL=AVEABL/FLOAT(LIMIT)
C
C CALACULATE ROOT MEAN SQUARE FOR ABILITY LEVELS
C
      DO 1200 I=1,LIMIT
      DO 1100 J=1,ITEMS
      RMSLIM(I)=RMSLIM(I)+((RESID(I,J)-AVELIM(I))**2)
 1100 CONTINUE
      RMSLIM(I)=SQRT(RMSLIM(I)/FLOAT(ITEMS))
      AVERML=AVERML+RMSLIM(I)
 1200 CONTINUE
      AVERML=AVERML/FLOAT(LIMIT)
C
C CALACULATE FOR AVERAGE ITEM STATISTIC , AVERAGE ABSOLUTE ITEM STATISTICS
C
      DO 1400 I=1,ITEMS
      DO 1300 J=1,LIMIT
      AVEITM(I)=AVEITM(I)+RESID(J,I)
      ABSITM(I)=ABSITM(I)+ABS(RESID(J,I))
 1300 CONTINUE
      AVEITM(I)=AVEITM(I)/FLOAT(LIMIT)
      ABSITM(I)=ABSITM(I)/FLOAT(LIMIT)
      AVEAVI=AVEAVI+AVEITM(I)
      AVEABI=AVEABI+ABSITM(I)
 1400 CONTINUE
      AVEAVI=AVEAVI/FLOAT(ITEMS)
      AVEABI=AVEABI/FLOAT(ITEMS)
C
C CALCULATE ROOT MEAN SQUARE FOR ITEM STATISTICS
C
      DO 1600 I=1,ITEMS
      DO 1500 J=1,LIMIT
      RMSITM(I)=RMSITM(I)+((RESID(J,I)-AVEITM(I))**2)
 1500 CONTINUE
      RMSITM(I)=SQRT(RMSITM(I)/FLOAT(LIMIT))
      AVERMI=AVERMI+RMSITM(I)
 1600 CONTINUE
      AVERMI=AVERMI/FLOAT(ITEMS)
C
C CALCULATE THE WEIGHTED AVERAGE RESIDUALS AND
C         WEIGHTED AVERAGE ABSOLUTE RESIDUALS
C
      KTOTAL=NSUBJ-KCOUNT
      DO 1650 I=1,ITEMS
      DO 1640 J=1,LIMIT
      WAVITM(I)=WAVITM(I)+(RESID(J,I)*FLOAT(IR(J,ITEMONE)))
      WABITM(I)=WABITM(I)+(ABS(RESID(J,I))*FLOAT(IR(J,ITEMONE)))
 1640 CONTINUE
      WAVITM(I)=WAVITM(I)/FLOAT(KTOTAL)
      WABITM(I)=WABITM(I)/FLOAT(KTOTAL)
      AVEWAV=AVEWAV+WAVITM(I)
      AVEWAB=AVEWAB+WABITM(I)
 1650 CONTINUE
      AVEWAV=AVEWAV/FLOAT(ITEMS)
      AVEWAB=AVEWAB/FLOAT(ITEMS)
C
C PRINT OUT FIT STATISTICS FOR ITEMS
C
      WRITE (8,1700)
 1700 FORMAT(5X,'SUMMARY OF FIT STATISTICS FOR ITEMS',//,
     *5X,'ITEM  AVERAGE  AVERAGE ABSOLUTE  ROOT MEAN',8X,
     *'WEIGHTED AVERAGE  WEIGHTED ABSOLUTE',/,
     *5X,'     RESIDUAL    RESIDUAL          SQUARE RESIDUAL ',6X,
     *'RESIDUAL      AVERAGE  RESIDUAL')
      DO 1715 I=1,ITEMS
      WRITE(8,1710) I,AVEITM(I),ABSITM(I),RMSITM(I),WAVITM(I),WABITM(I)
 1710 FORMAT(6X,I2,2X,F7.3,5X,F7.3,9X,F7.3,14X,F7.3,11X,F7.3)
      K=MOD(I,L)
      IF(K.NE.0)GO TO 1705
      WRITE(8,1702)
 1702 FORMAT(/)
 1705 CONTINUE
 1715 CONTINUE
C
C PRINT AVERAGES FOR ITEM FIT STATISTICS
C
      WRITE (8,1720) AVEAVI,AVEABI,AVERMI,AVEWAV,AVEWAB
 1720 FORMAT(/,2X,'AVERAGES',F7.3,5X,F7.3,9X,F7.3,
     *14X,F7.3,11X,F7.3)
C
C PRINT ABILITY LEVEL FIT STATISTICS
C
      WRITE(8,471)
      WRITE(8,1721)
 1721 FORMAT(40X,'SUMMARY OF FIT STATISTICS FOR ABILITY LEVELS',//,
     *50X,'ABILITY LEVEL(MID-POINTS)',/)
      WRITE(8,1750)(I,I=1,LIMIT)
      WRITE(8,1722) (ABIL(J),J=1,LIMIT)
 1722 FORMAT(1X,'FIT',1X,'STATISTIC',15X,15(F7.2,1X))
 1750 FORMAT(25X,15(5X,I3),/)
      WRITE(8,1760) (AVELIM(I),I=1,LIMIT)
 1760 FORMAT(1X,'AVERAGE RESIDUAL',12X,15(F7.3,1X))
      WRITE(8,1770) (ABSLIM(I),I=1,LIMIT)
 1770 FORMAT(//,1X,'AVERAGE ABSOLUTE RESIDUAL',2X,15(F7.3,1X))
      WRITE (8,1780) (RMSLIM(I),I=1,LIMIT)
 1780 FORMAT(//,1X,'ROOT MEAN SQUARE RESIDUAL',2X,15(F7.3,1X))
C
C PRINT AVERAGES FOR ABILITY LEVEL FIT STATISTICS
C
      WRITE (8,471)
      WRITE (8,1790) AVEAVL,AVEABL,AVERML
  471 FORMAT(////)
 1790 FORMAT(5X,'OVERALL  VERAGES   AVERAGE RESIDUAL= ',F7.3,3X,
     *'AVERAGE ABSOLUTE RESIDUAL= ',F7.3,3X,'ROOT MEAN SQUARE',
     *' RESIDUAL= ',F7.3)
      END
```

```
RESIDUAL ANALYSES OF LOGISTIC TEST DATA
              DATE: 82/04/09.
              (VERSION 3A)


      PROGRAM BY LINDA MURRAY
                 RONALD HAMBLETON
                 ROBERT SIMON


DEVELOPED AT THE UNIVERSITY OF MASSACHUSETTS
        SCHOOL OF EDUCATION
     UNDER A GRANT FROM NAEP


NAEP DATA-MATH RESULTS OF 13 YRS. OLD BOOK 1



 58 ITEMS /  2422 EXAMINEES / 12 ABILITY GROUPINGS
```

EXPECTED P VALUES-1 PARAMETER MODEL

	ABILITY LEVEL											
CATEGORY	1	2	3	4	5	6	7	8	9	10	11	12
MID-POINT	-2.75	-2.25	-1.75	-1.25	-.75	-.25	.25	.75	1.25	1.75	2.25	2.75
NO. OF EXAMINEES	14	54	91	224	325	503	467	339	245	102	44	3
ITEM												
1	.243	.409	.548	.681	.789	.868	.920	.953	.973	.984	.990	.990
2	.489	.627	.747	.838	.901	.941	.965	.980	.989	.990	.990	.990
3	.621	.742	.835	.899	.940	.965	.980	.988	.990	.990	.990	.990
4	.048	.081	.134	.213	.322	.455	.594	.720	.819	.888	.933	.961
5	.087	.144	.228	.341	.476	.615	.737	.831	.896	.939	.964	.979
6	.021	.036	.062	.104	.170	.264	.387	.525	.660	.773	.857	.913
7	.027	.046	.078	.129	.206	.313	.445	.584	.712	.813	.884	.930
8	.115	.185	.286	.412	.552	.684	.791	.869	.921	.954	.973	.984
9	.015	.027	.046	.078	.129	.206	.313	.445	.584	.712	.813	.884
10	.214	.314	.457	.596	.722	.820	.889	.933	.961	.977	.987	.990
11	.628	.748	.839	.901	.941	.966	.980	.989	.990	.990	.990	.990
12	.538	.671	.782	.863	.917	.951	.971	.984	.990	.990	.990	.990
13	.490	.628	.748	.839	.901	.941	.966	.980	.989	.990	.990	.990
14	.471	.610	.733	.828	.894	.937	.963	.979	.988	.990	.990	.990
15	.369	.507	.644	.760	.848	.907	.945	.968	.981	.989	.990	.990
16	.350	.496	.624	.744	.836	.900	.940	.965	.980	.988	.990	.990
17	.132	.211	.320	.452	.592	.718	.817	.887	.932	.960	.977	.987
18	.010	.010	.014	.025	.043	.072	.121	.194	.297	.426	.566	.696
19	.077	.127	.204	.310	.441	.580	.708	.816	.882	.929	.958	.976
20	.069	.113	.183	.282	.428	.548	.680	.759	.858	.920	.953	.973
21	.056	.095	.156	.245	.362	.499	.637	.755	.844	.905	.943	.967
22	.526	.661	.774	.857	.913	.949	.970	.983	.990	.990	.990	.990
23	.022	.037	.064	.107	.174	.270	.393	.532	.666	.778	.866	.915
24	.019	.031	.053	.090	.148	.233	.348	.484	.622	.743	.835	.899
25	.047	.079	.131	.209	.317	.448	.588	.715	.815	.885	.931	.960
26	.135	.215	.325	.458	.597	.722	.820	.889	.934	.961	.977	.987
27	.010	.010	.010	.017	.030	.051	.087	.143	.227	.340	.475	.613
28	.010	.015	.027	.047	.079	.131	.210	.318	.450	.589	.716	.816
29	.010	.012	.021	.037	.063	.106	.172	.267	.390	.528	.663	.775
30	.035	.059	.100	.163	.254	.374	.512	.648	.764	.850	.909	.946
31	.146	.231	.345	.481	.619	.741	.834	.898	.939	.964	.979	.988
32	.062	.104	.169	.263	.385	.524	.659	.772	.856	.913	.948	.970
33	.029	.051	.085	.141	.224	.336	.470	.609	.732	.828	.894	.937
34	.636	.754	.843	.904	.943	.967	.981	.989	.990	.990	.990	.990
35	.095	.156	.245	.363	.500	.637	.755	.844	.905	.943	.967	.981
36	.010	.016	.027	.046	.079	.130	.208	.316	.448	.587	.714	.814
37	.022	.038	.064	.108	.175	.271	.395	.534	.668	.779	.861	.916
38	.180	.279	.404	.543	.676	.786	.866	.919	.952	.972	.984	.990
39	.114	.184	.283	.410	.549	.682	.790	.868	.920	.953	.973	.984
40	.092	.151	.238	.354	.490	.628	.747	.839	.901	.941	.966	.980
41	.013	.022	.038	.065	.108	.176	.272	.396	.535	.669	.780	.862
42	.109	.177	.274	.399	.539	.671	.782	.863	.917	.951	.971	.984
43	.178	.275	.400	.539	.672	.783	.863	.917	.951	.972	.984	.990
44	.103	.167	.261	.382	.521	.656	.770	.855	.912	.949	.970	.982
45	.034	.058	.097	.159	.248	.367	.505	.641	.758	.846	.906	.944
46	.019	.033	.057	.096	.157	.246	.364	.502	.639	.756	.845	.905
47	.010	.010	.011	.019	.033	.057	.096	.158	.247	.366	.503	.640
48	.010	.010	.016	.027	.047	.080	.132	.211	.320	.452	.592	.718
49	.041	.071	.118	.190	.291	.419	.559	.690	.796	.873	.923	.955
50	.260	.382	.520	.655	.769	.854	.911	.948	.969	.982	.990	.990
51	.010	.012	.021	.036	.061	.103	.168	.262	.383	.522	.657	.771
52	.071	.118	.190	.291	.419	.559	.690	.796	.873	.923	.955	.974
53	.233	.344	.483	.618	.740	.833	.898	.939	.964	.979	.988	.990
54	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.933	.961	.977	.987

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
29	[illegible]	[illegible]	[illegible]	.037	.063	.116	.172	.267	.391	.528	.663	.775
30	.035	.059	[illegible]	.163	.254	.374	.512	.648	.764	.850	.909	.946
31	.146	.231	.345	.481	.619	.741	.834	.898	.939	.964	.979	.988
32	.052	.094	.149	.263	.385	.526	.659	.772	.856	.913	.948	.970
33	.029	.061	.095	.141	.224	.336	.470	.609	.732	.828	.894	.937
34	.636	.754	.843	.904	.943	.967	.981	.989	.993	.990	.996	.990
35	.195	.255	.245	.363	.500	.627	.755	.844	.905	.943	.967	.981
36	.010	.016	.027	.046	.079	.130	.208	.316	.448	.587	.714	.814
37	.022	.039	.064	.108	.175	.271	.395	.534	.668	.779	.861	.916
38	.180	.279	.414	.543	.676	.786	.866	.919	.952	.972	.984	.990
39	.124	.194	.283	.420	.549	.682	.790	.868	.920	.953	.973	.984
40	.092	.151	.238	.354	.490	.628	.747	.839	.901	.941	.966	.980
41	.013	.022	.038	.065	.108	.176	.272	.396	.535	.669	.780	.862
42	.109	.177	.274	.399	.533	.671	.782	.863	.917	.951	.971	.984
43	.178	.265	.410	.539	.672	.783	.863	.917	.951	.972	.984	.990
44	.103	.167	.261	.382	.521	.656	.770	.855	.912	.949	.970	.982
45	.034	.058	.097	.153	.248	.367	.505	.641	.758	.846	.906	.944
46	.019	.033	.057	.096	.157	.246	.364	.502	.639	.756	.845	.905
47	.013	[illegible]	.021	.019	.033	.057	.096	.158	.247	.366	.503	.640
48	.013	.013	.016	.027	.047	.080	.132	.211	.320	.452	.592	.718
49	.041	.071	.118	.190	.291	.419	.559	.690	.796	.873	.923	.955
50	.260	.372	.520	.655	.769	.854	.921	.948	.969	.982	.990	.990
51	.010	.012	.021	.036	.061	.103	.168	.252	.383	.522	.657	.771
52	.071	.118	.190	.291	.419	.559	.690	.796	.873	.923	.955	.974
53	.230	.344	.490	.628	.740	.833	.898	.939	.964	.979	.988	.990
54	.134	.214	.324	.457	.596	.721	.820	.889	.933	.961	.977	.987
55	.011	.019	.033	.057	.096	.158	.247	.366	.502	.640	.757	.845
56	.129	.206	.313	.444	.584	.711	.812	.881	.930	.959	.976	.986
57	.015	.029	.048	.082	.135	.215	.325	.458	.597	.722	.820	.889
58	.147	.233	.347	.483	.621	.742	.835	.899	.940	.965	.980	.988

OBSERVED P VALUES-1 PARAMETER MODEL

```
                                                  ABILITY LEVEL
CATEGORY       1      2      3      4      5      6      7      8      9     10     11     12

MID-POINT   -2.75  -2.25  -1.75  -1.25   -.75   -.25    .25    .75   1.25   1.75   2.25   2.75
NO. OF
EXAMINEES      14     54     91    224    325    503    467    339    245    102     44      3

  ITEM
   1         .286   .204   .418   .612   .778   .907   .942   .982   .971  1.000   .977  1.000
   2         .429   .630   .604   .844   .902   .952   .974   .985   .988  1.000  1.000  1.000
   3         .643   .685   .758   .893   .945   .970   .991   .994   .992  1.000  1.000  1.000
   4        0.000   .056   .044   .076   .178   .380   .655   .870   .967   .980   .977  1.000
   5         .071   .093   .121   .268   .375   .604   .797   .903   .939   .971   .977  1.000
   6         .071   .111   .220   .183   .188   .235   .351   .496   .653   .775   .932  1.000
   7         .143   .111   .176   .201   .225   .302   .396   .543   .706   .853   .977  1.000
   8         .571   .556   .549   .585   .643   .636   .707   .767   .845   .912   .977  1.000
   9        0.000   .185   .121   .103   .145   .268   .296   .375   .522   .637   .841   .667
  10         .214   .222   .319   .491   .698   .829   .934   .976   .988   .990  1.000  1.000

  11         .071   .556   .802   .888   .963   .982   .996   .994   .996  1.000  1.000  1.000
  12         .143   .407   .692   .866   .938   .972   .985   .994   .992  1.000  1.000  1.000
  13         .071   .426   .626   .813   .917   .968   .994   .988   .988  1.000  1.000  1.000
  14         .071   .444   .703   .808   .898   .948   .979   .994   .996  1.000  1.000  1.000
  15         .143   .389   .538   .750   .855   .915   .974   .968   .984  1.000  1.000  1.000
  16         .071   .352   .593   .763   .846   .930   .938   .947   .967  1.000  1.000  1.000
  17         .429   .333   .440   .455   .594   .771   .760   .861   .890   .902  1.000  1.000
  18        0.000  0.000  0.000   .009   .031   .062   .146   .189   .314   .480   .477  1.000
  19         .071   .019   .044   .156   .323   .642   .771   .855   .931   .971  1.000  1.000
  20         .071   .019   .033   .107   .215   .535   .777   .941   .955  1.000  1.000  1.000

  21        0.000   .019   .044   .067   .200   .489   .700   .894   .947  1.000  1.000  1.000
  22         .571   .704   .868   .853   .905   .946   .957   .985   .984   .990  1.000  1.000
  23         .143   .148   .165   .268   .252   .272   .334   .428   .600   .755   .932  1.000
  24        0.000   .130   .187   .170   .151   .219   .323   .442   .584   .784   .932  1.000
  25         .071   .019   .121   .112   .268   .475   .623   .717   .849   .912   .955  1.000
  26         .214   .593   .549   .513   .643   .672   .762   .873   .878   .961  1.000  1.000
  27         .071  0.000   .011   .013   .034   .054   .086   .112   .245   .314   .636  1.000
  28         .071   .130   .143   .152   .175   .195   .238   .245   .233   .382   .364   .667
  29        0.000   .019   .011   .022   .068   .091   .135   .301   .441   .529   .727  1.000
  30        0.000   .130   .187   .210   .157   .328   .527   .664   .820   .912  1.000   .667

  31         .214   .241   .330   .415   .594   .738   .859   .917   .947  1.000   .977  1.000
  32         .143   .204   .242   .228   .326   .459   .657   .835   .939   .980   .977   .667
  33        0.000   .093   .165   .165   .283   .338   .454   .611   .637   .814   .886  1.000
  34         .714   .759   .830   .884   .960   .966   .974   .976   .996  1.000  1.000  1.000
  35         .429   .426   .440   .455   .526   .579   .690   .825   .910   .961  1.000  1.000
  36         .143   .148   .187   .232   .262   .215   .218   .171   .208   .196   .318  0.000
  37        0.000   .093   .165   .161   .212   .266   .375   .519   .624   .745   .818  1.000
  38        0.000   .074   .121   .433   .618   .839   .919   .971   .992  1.000  1.000  1.000
  39         .429   .315   .352   .531   .625   .716   .747   .773   .833   .863   .886   .667
  40         .071   .037   .132   .281   .443   .606   .799   .900   .931   .990  1.000  1.000



  41        0.000   .037   .044   .063   .065   .147   .272   .381   .630   .814   .886  1.000
  42         .214   .204   .264   .388   .520   .660   .782   .870   .959   .941   .977  1.000
  43         .286   .370   .319   .513   .646   .777   .874   .941   .963   .980   .977  1.000
  44        0.000   .037   .154   .371   .526   .712   .794   .850   .873   .892   .886  1.000
  45        0.000  0.000   .044   .098   .222   .414   .529   .631   .780   .833   .864  1.000
  46         .143   .074   .143   .080   .112   .183   .358   .575   .678   .833   .955  1.000
  47         .071   .093   .066   .116   .128   .091   .122   .109   .082   .225   .227   .667
  48         .143   .056   .066   .040   .040   .066   .126   .201   .335   .480   .545   .667
  49        0.000   .093   .033   .121   .249   .394   .604   .743   .820   .902   .909  1.000
  50         .143   .315   .473   .603   .803   .855   .916   .950   .988   .980  1.000  1.000

  51         .071   .093   .121   .107   .129   .137   .163   .206   .253   .412   .568  1.000
  52         .286   .463   .418   .473   .517   .575   .638   .687   .755   .735   .750   .333
  53         .357   .463   .550   .665   .769   .809   .878   .903   .955   .951  1.000  1.000
  54        0.000   .074   .198   .438   .578   .765   .844   .888   .922   .941  1.000  1.000
```

```
26        .216  .593  .549  .513  .643  .672  .762  .873  .878  .961 1.000 1.000
27        .071 0.000  .011  .023  .034  .054  .086  .112  .245  .314  .636 1.000
28        .071  .130  .143  .152  .175  .195  .238  .245  .233  .382  .364  .667
29       0.000  .019  .011  .022  .068  .091  .135  .301  .441  .529  .727 1.000
30       0.000  .130  .137  .210  .157  .328  .527  .664  .820  .912 1.000  .667

31        .214  .241  .330  .415  .594  .738  .859  .917  .947 1.000  .977 1.000
32        .143  .204  .242  .226  .326  .459  .657  .835  .939  .980  .977  .667
33       0.000  .093  .165  .155  .283  .338  .454  .611  .637  .814  .886 1.000
34        .714  .759  .830  .884  .960  .966  .974  .976  .996 1.000 1.000 1.000
35        .429  .426  .440  .455  .526  .579  .699  .825  .910  .961 1.000 1.000
36        .143  .148  .187  .232  .262  .215  .218  .171  .208  .196  .318 0.000
37       0.000  .093  .165  .161  .212  .266  .375  .519  .624  .745  .818 1.000
38       0.000  .074  .121  .433  .618  .839  .919  .971  .992 1.000 1.000 1.000
39        .429  .315  .352  .531  .625  .716  .747  .773  .833  .863  .886  .667
40        .071  .037  .132  .281  .443  .606  .799  .960  .931  .990 1.000 1.000


41       0.000  .037  .044  .063  .065  .147  .272  .381  .600  .814  .886 1.000
42        .214  .204  .264  .388  .520  .660  .782  .870  .959  .941  .977 1.000
43        .286  .370  .319  .513  .646  .777  .874  .941  .963  .980  .977 1.000
44       0.000  .037  .154  .371  .526  .712  .794  .850  .873  .892  .886 1.000
45       0.000 0.000  .044  .098  .222  .414  .529  .631  .786  .833  .864 1.000
46        .143  .074  .143  .080  .132  .183  .358  .575  .678  .833  .955 1.000
47        .071  .093  .066  .116  .108  .091  .122  .109  .082  .225  .227  .667
48        .143  .056  .066  .044  .040  .066  .126  .201  .335  .480  .545  .667
49       0.000  .093  .088  .121  .249  .394  .604  .743  .820  .902  .909 1.000
50        .143  .315  .473  .603  .803  .855  .916  .950  .988  .980 1.000 1.000

51        .071  .093  .121  .107  .129  .137  .163  .206  .253  .412  .568 1.000
52        .286  .413  .418  .473  .517  .575  .638  .687  .755  .735  .750  .333
53        .357  .463  .560  .665  .769  .809  .878  .903  .955  .951 1.000 1.000
54       0.000  .074  .198  .438  .578  .765  .844  .888  .922  .941 1.000 1.000
55       0.000  .019  .011  .013  .034  .117  .229  .445  .584  .735  .886 1.000
56        .286  .167  .253  .433  .582  .716  .814  .900  .927  .961  .977 1.000
57        .071  .019  .011  .013  .037  .141  .360  .555  .706  .824  .909 1.000
58        .286  .148  .297  .451  .609  .750  .854  .897  .947  .990 1.000 1.000
```

```
                                   RESIDUALS- 1 PARAMETER MODEL

                                      (OBSERVED-EXPECTED)

                                                   ABILITY LEVEL
                 1       2       3       4       5       6       7       8       9      10      11      12

MID-POINT    -2.750  -2.250  -1.750  -1.250   -.750   -.250    .250    .750   1.250   1.750   2.250   2.750
NO. OF
EXAMINEES       14      54      91     224     325     503     467     339     245     102      44       3

  ITEM
    1         .003   -.205   -.131   -.069   -.011    .039    .022    .029   -.001    .016   -.013    .010
    2        -.060    .003   -.142    .006    .001    .011    .009    .005   -.001    .010    .010    .010
    3         .022   -.057   -.077   -.046    .005    .005    .012    .006    .002    .010    .010    .010
    4        -.048   -.025   -.090   -.137   -.144   -.075    .061    .150    .149    .092    .044    .035
    5        -.016   -.051   -.107   -.073   -.101   -.010    .060    .072    .043    .033    .014    .021
    6         .050    .075    .158    .079    .018   -.030   -.035   -.030   -.007    .001    .075    .087
    7         .116    .065    .098    .072    .018   -.011   -.049   -.042   -.006    .040    .093    .070
    8         .457    .376    .264    .173    .091   -.046   -.085   -.10[illegible]   -.076   -.042    .064    .016
    9        -.015    .259    .075    .025    .016    .062   -.018   -.070   -.062   -.074    .028   -.217
   10        -.000   -.102   -.138   -.105   -.023    .009    .045    .043    .027    .013    .013    .010

   11        -.557   -.192   -.037   -.013    .022    .016    .016    .005    .006    .010    .010    .010
   12        -.395   -.264   -.089    .003    .022    .021    .014    .011    .002    .010    .010    .010
   13        -.419   -.202   -.121   -.026    .016    .027    .028    .008   -.001    .010    .010    .010
   14        -.399   -.165   -.029   -.020    .004    .011    .016    .016    .008    .010    .010    .010
   15        -.227   -.119   -.105   -.010    .008    .007    .029   -.000    .002    .011    .010    .010
   16        -.278   -.134   -.036    .019    .010    .031   -.002   -.018   -.012    .012    .010    .010
   17         .296    .122    .120    .003    .002    .054   -.057   -.025   -.042   -.058    .023    .013
   18        -.010   -.010   -.014   -.016   -.012   -.011    .025   -.005    .017    .054   -.088    .304
   19        -.035   -.109   -.160   -.154   -.118    .062    .063    .046    .049    .041    .042    .024
   20         .004   -.095   -.150   -.175   -.193   -.013    .097    .152    .087    .080    .047    .027

   21        -.056   -.077   -.112   -.178   -.162   -.010    .064    .139    .103    .095    .057    .033
   22         .045    .043    .094   -.005   -.009   -.002   -.013    .002   -.036    .000    .010    .010
   23         .121    .111    .101    .161    .079    .003   -.059   -.104   -.066   -.023    .072    .085
   24        -.018    .099    .134    .086    .003   -.014   -.025   -.041   -.038    .042    .097    .101
   25         .025   -.060   -.010   -.097   -.049    .027    .035    .002    .034    .026    .023    .040
   26         .151    .377    .225    .055    .046   -.050   -.058   -.016   -.056   -.000    .023    .013
   27         .061   -.010    .051   -.004    .004    .002   -.001   -.031    .018   -.026    .162    .387
   28         .051    .114    .116    .105    .096    .064    .028   -.073   -.217   -.207   -.352   -.149
   29        -.010    .036   -.010   -.015    .005   -.014   -.037    .034    .051    .001    .064    .225
   30        -.035    .070    .087    .047   -.097   -.046    .014    .015    .056    .061    .091   -.279

   31         .058    .010   -.016   -.066   -.025   -.003    .025    .019    .008    .036   -.002    .012
   32         .081    .100    .073   -.035   -.059   -.065   -.042    .063    .083    .068    .029   -.303
   33        -.029    .042    .079    .024    .059    .002   -.016    .002   -.096   -.014   -.010    .063
   34         .079    .005    .047   -.020    .017   -.001   -.006   -.013    .006    .010    .010    .010
   35         .333    .270    .195    .093    .026   -.058   -.065   -.039    .006    .017    .033    .019
   36         .133    .133    .160    .186    .183    .084    .010   -.145   -.239   -.391   -.396   -.814
   37        -.022    .055    .100    .053    .037   -.005   -.020   -.015   -.044   -.034   -.043    .084
   38        -.180   -.205   -.283   -.110   -.058    .053    .053    .052    .040    .028    .016    .010
   39         .315    .131    .068    .121    .075    .034   -.042   -.095   -.088   -.090   -.086   -.318
   40        -.020   -.114   -.106   -.072   -.047   -.021    .051    .061    .029    .049    .034    .020

   41        -.013    .015    .006   -.002   -.044   -.029   -.000   -.016    .065    .145    .105    .138


   42         .105    .027   -.010   -.010   -.018   -.011   -.000    .007    .042   -.010    .006    .016
   43         .128    .095   -.081   -.026   -.026   -.005    .010    .024    .012    .009   -.066    .010
   44        -.103   -.110   -.107   -.012    .005    .056    .024   -.005   -.038   -.056   -.083    .016
   45        -.034   -.153   -.053   -.060   -.027    .046    .024   -.016    .021   -.013   -.043    .056
   46         .124    .341    .096   -.016   -.055   -.063   -.007    .074    .039    .077    .110    .095
   47         .061    .383    .455    .097    .074    .034    .026   -.048   -.165   -.140   -.276    .327
```

```
42          .135   .027  -.010  -.012  -.018  -.011  -.006   .007   .042  -.010   .006   .018

43          .133   .095  -.081  -.026  -.026  -.005   .010   .024   .012   .009  -.066   .010
44         -.103  -.110  -.107  -.012   .005   .056   .024  -.005  -.038  -.056  -.083   .018
45         -.034  -.159  -.053  -.060  -.027   .046   .024  -.010   .021  -.013  -.043   .058
46          .124   .041   .096  -.016  -.055  -.063  -.007   .074   .039   .077   .110   .095
47          .061   .083   .055   .097   .074   .034   .026  -.048  -.165  -.140  -.276   .327
48          .133   .045   .056   .013  -.007  -.014  -.006  -.011   .015   .028  -.046  -.051
49         -.041   .022  -.030  -.069  -.042  -.025   .045   .054   .024   .029  -.014   .045
50         -.117  -.067  -.047  -.053   .034   .001   .005   .002   .018  -.002   .010   .010

51          .061   .081   .100   .071   .068   .034  -.005  -.055  -.130  -.110  -.089   .229
52          .215   .345   .228   .182   .098   .016  -.052  -.149  -.118  -.188  -.205  -.640
53          .127   .119   .031   .047   .030  -.024  -.020  -.036  -.009  -.028   .012   .010
54         -.134  -.040  -.126  -.019  -.017   .044   .024  -.001  -.011  -.020   .023   .013
55         -.011  -.001  -.022  -.044  -.062  -.040  -.018   .080   .081   .096   .129   .155
56          .157  -.039  -.060  -.011  -.002   .065   .002   .016  -.004   .002   .001   .014
57          .055  -.010  -.037  -.068  -.098  -.074   .035   .097   .139   .101   .089   .111
58          .139  -.084  -.051  -.032  -.012   .007   .026  -.002   .007   .026   .020   .012


SUMMARY OF FIT STATISTICS FOR ITEMS

ITEM  AVERAGE   AVERAGE ABSOLUTE  ROOT MEAN         WEIGHTED AVERAGE   WEIGHTED ABSOLUTE
      RESIDUAL      RESIDUAL      SQUARE RESIDUAL       RESIDUAL        AVERAGE   RESIDUAL
  1    -.026          .046            .070               -.001              .035
  2    -.012          .022            .044                .000              .012
  3    -.005          .018            .029                .001              .011
  4     .001          .088            .099                .001              .105
  5    -.011          .050            .058               -.000              .055
  6     .037          .054            .057                .001              .037
  7     .039          .057            .053                .000              .037
  8     .085          .144            .181               -.001              .099
  9    -.009          .068            .090                .001              .047
 10    -.017          .044            .059               -.000              .041
 11    -.059          .075            .160                .001              .021
 12    -.054          .071            .129                .001              .024
 13    -.055          .073            .128                .001              .029
 14    -.044          .058            .118                .001              .019
 15    -.032          .045            .074                .000              .018
 16    -.032          .047            .085                .000              .020
 17     .039          .068            .097               -.000              .043
 18    [illegible]    .047            .092                .001              .018
 19    -.018          .073            .085                .000              .078
 20    -.011          .093            .111                .001              .106
 21    -.019          .091            .103                .001              .094
 22     .014          [illegible]     .030                .000              .011
 23    [illegible]    .082            .083                .001              .068
 24     .015          .058            .061                .001              .036
 25    [illegible]    .036            .043                .001              .035
 26     .059          .089            .125               -.000              .059
 27     .047          .059            .114                .000              .013
 28    -.035          .132            .153                .001              .096
 29     .025          .039            .066                .000              .024
 30    -.031          .075            .101                .001              .048
 31     .015          .024            .032               -.000              .021
 32     .003          .080            .108                .001              .051
 33    [illegible]    .036            .046                .001              .029
 34     .012          .019            .026                .001              .011
 35     .059          .096            .124               -.001              .060
 36    -.031          .239            .361                .002              .141
 37     .012          .043          .049               .001          .029
 38    -.009          .091          .113              -.000          .069
 39     .002          .122          .152              -.000          .072
 40    -.011          .052          .059               .000          .048
 41     .031          .048          .063               .001          .030
 42     .012          .022          .033              -.000          .013
 43     .010          .034          .049              -.001          .020
 44    -.036          .053          .057              -.000          .034
 45    -.012          .037          .039               .001          .032
 46     .042          .065          .062               .001          .048
 47    -.014          .091          .113               .001          .070
 48     .012          .035          .047               .000          .015
 49    -.0[illegible]          .037          .040               .001          .040
 50    -.017          .031          .042              -.001          .017
 51     .021          .086          .099               .001          .057
 52    -.019          .20[illegible]          .254               .001          .101
 53     .026          .045          .054              -.001          .031
 54    -.030          .048          .062              -.001          .029
 55     .028          .062          .071               .001          .051
 56     .007          .026          .050              -.000          .009
 57     .026          .074          .076               .001          .074
 58     .004          .034          .051              -.000          .017

AVERAGES .001         .065          .086               .000          .044


                                  SUMMARY OF FIT STATISTICS FOR ABILITY LEVELS
                                         ABILITY LEVEL(MID-POINTS)
                              1        2        3        4        5        6        7        8        9       10       11       12
FIT STATISTIC             -2.75    -2.25    -1.75    -1.25     -.75     -.25      .25      .75     1.25     1.75     2.25     2.75
AVERAGE RESIDUAL           .008     .009     .004    -.002    -.006     .002     .005     .002    -.005    -.002    -.001    -.000

AVERAGE ABSOLUTE RESIDUAL  .120     .103     .093     .061     .046     .028     .029     .042     .048     .050     .059     .096

ROOT MEAN SQUARE RESIDUAL  .176     .135     .112     .081     .065     .036     .036     .059     .071     .082     .099     .180


  OVERALL AVERAGES   AVERAGE RESIDUAL=    .001   AVERAGE ABSOLUTE RESIDUAL=    .065   ROOT MEAN SQUARE RESIDUAL=    .094
```

```
                             STANDARDIZED RESIDUALS- 1 PARAMETER MODEL

                                                          ABILITY LEVEL
                1       2       3       4       5       6       7       8       9      10      11      12

MID-POINT   -2.750  -2.250  -1.750  -1.250   -.750   -.250    .250    .750   1.250   1.750   2.250   2.750
NO. OF
EXAMINEES      14      54      91     224     325     563     467     339     245     102      44       3

  ITEM
    1         .026  -3.665  -2.505  -2.213   -.466   2.566   1.757   2.556   -.113   1.279   -.848    .174
    2        -.456    .846  -3.118    .235    .046   1.679   1.044    .686   -.114   1.015    .667    .174
    3         .166   -.959  -1.959   -.292    .373    .663   1.810    .997    .289   1.015    .667    .174
    4        -.237   -.683  -2.514  -5.014  -5.543  -3.383   2.685   6.161   6.642   2.961   1.176    .351
    5        -.211  -1.073  -2.431  -2.313  -3.636   -.475   2.930   3.522   2.187   1.363    .480    .254
    6        1.311   2.928   6.219   3.853    .857  -1.507  -1.574  -1.096   -.234    .031   1.420    .534
    7        2.700   2.292   3.491   3.213    .814   -.541  -2.116  -1.556   -.193   1.046   1.935    .474
    8        5.361   6.998   5.574   5.246   3.307  -2.197  -4.510  -5.603  -4.434  -2.005    .175    .218
    9        -.467   7.233   3.421   1.390    .840   3.439   -.832  -2.600  -1.969  -1.660    .483  -1.174
   10        -.001  -1.596  -2.645  -3.265   -.928    .540   3.687   3.175   2.166    .871    .762    .174

   11       -4.313  -3.258   -.956   -.657   1.665   2.018   2.408    .951    .931   1.015    .667    .174
   12       -2.963  -4.126  -2.067    .141   1.405   2.205   1.760   1.529    .289   1.015    .667    .174
   13       -3.136  -3.075  -2.668  -1.073    .340   2.564   3.309   1.066   -.126   1.015    .667    .174
   14       -2.994  -2.488   -.634   -.789    .253   1.659   1.784   1.974   1.168   1.015    .667    .174
   15       -1.757  -1.737  -2.093   -.356    .388    .571   2.786   -.030    .261   1.049    .667    .174
   16       -2.184  -1.967   -.597    .658    .482   2.296   -.215  -1.822  -1.385   1.095    .667    .174
   17        3.271   2.199   2.452    .100    .085   2.675  -3.171  -1.478  -2.640  -3.010   1.019    .201
   18        -.376   -.739  -1.147  -1.523  -1.058   -.939   1.656   -.247    .587   1.111  -1.184   1.149
   19        -.673  -2.396  -3.792  -4.569  -4.269   2.809   2.978   2.137   2.356   1.629   1.382    .272
   20         .054  -2.196  -3.703  -5.825  -7.082   -.594   4.495   6.865   4.039   2.977   1.476    .291

   21        -.915  -1.918  -2.941  -6.185  -6.091   -.465   2.860   5.957   4.451   3.281   1.627    .321
   22         .339    .664   2.150   -.196   -.561   -.246  -1.646    .351   -.395    .020    .667    .174
   23        3.116   4.289   3.940   7.794   3.742    .144  -2.612  -3.849  -2.197   -.560   1.371    .527
   24        -.565   4.179   5.678   4.186    .160   -.766  -1.117  -1.518  -1.232    .961   1.728    .581
   25         .443  -1.145   -.277  -3.577  -1.893   1.206   1.543    .086   1.380    .839    .610    .355
   26        1.649   6.750   4.575   1.667   1.686  -2.528  -3.272   -.937  -3.531   -.017   1.807    .198
   27        2.310   -.739    .095   -.447    .411    .229   -.096  -1.633    .678   -.557   2.147   1.375
   28        2.310   6.743   6.786   7.446   6.413   4.219   1.487  -2.883  -6.833  -4.249  -5.180   -.665
   29        -.376    .417   -.684  -1.156    .349  -1.028  -2.163   1.426   1.644    .021    .963    .932
   30        -.749   2.188   2.777   1.913  -4.031  -2.151    .622    .589   2.078   1.738   2.160  -2.140

   31         .721    .167   -.316  -1.968   -.941   -.153   1.453   1.185    .509   1.940   -.160    .190
   32        1.256   2.465   1.851  -1.265  -2.192  -2.903   -.069   2.746   3.685   2.423    .868  -3.073
   33        -.652   1.410   2.707   1.642   2.572    .102   -.764    .058  -3.377   -.371   -.163    .450
   34         .610    .091   1.236  -1.033   1.315   -.071  -1.022  -2.216    .911   1.015    .667    .174
   35        4.251   5.471   4.321   2.886    .353  -2.718  -3.279  -1.953    .297    .762   1.227    .242
   36        4.996   7.882   9.418  13.225  12.252   5.626    .544  -5.732  -7.536  -8.022  -5.810  -3.626
   37   [illegible]  2.116   3.904   2.554   1.772   -.246   -.906   -.557  -1.453   -.838   -.826    .525
   38       -1.755  -3.754  -5.506  -3.119  -2.231   2.906   3.360   3.488   2.916   1.111    .848    .174
   39        3.709   2.483   1.443   3.695   2.727   1.646  -2.250  -5.198  -5.081  -4.316  -3.517  -4.423
   40        -.265  -2.336  -2.370  -2.265  -1.688   -.988   2.551   3.060   1.544   2.103   1.251    .247
```

```
41      -.423   .757   .313  -.132 -2.531 -1.683  -.011  -.594  2.028  3.103  1.700   .694
42      1.261   .512  -.224  -.316  -.648  -.544  -.021   .368  2.392  -.459   .231   .224
43      1.057  1.569 -1.586  -.771 -1.009  -.295   .641  1.504   .877   .534  -.332   .174
44     -1.266 -2.565 -2.323  -.365   .193  2.626  1.251  -.265 -2.106 -2.519 -3.208   .232
45      -.658 -1.116 -1.706 -2.468 -1.119  2.160  1.055  -.382   .778  -.364  -.970   .420
46      3.369  1.672  3.515  -.789 -2.745 -3.298  -.309  2.768  1.268  1.814  2.010   .560
47      2.313  6.103  4.995 10.524  7.452  3.308  1.896 -2.445 -6.005 -2.937 -3.656   .097
48      4.596  3.365  3.834  1.169  -.509 -1.185  -.382  -.478   .503   .575  -.622  -.196
49      -.778   .630  -.882 -2.641 -1.659 -1.158  1.960  2.133   .947   .888  -.352   .377
50     -1.000 -1.016  -.906 -1.658  1.439   .042   .388   .193  1.667  -.150   .671   .174

51      2.310  5.451  6.687  5.720  5.086  2.512  -.300 -2.309 -4.198 -2.227 -1.243   .944
52      3.141  7.874  5.544  5.992  3.573   .711 -2.413 -4.968 -5.518 -7.130 -6.537 -6.935
53      1.127  1.833  1.541  1.449  1.214 -1.435 -1.393 -2.788  -.772 -2.012   .727   .174
54     -1.474 -2.510 -2.566  -.571  -.639  2.266  1.353  -.040  -.683 -1.028  1.010   .199
55      -.356  -.542 -1.190 -2.823 -3.815 -2.476  -.901  3.055  2.531  2.011  2.000   .741
56      1.756  -.711 -1.233  -.329  -.076   .232   .095   .932  -.220   .093   .047   .204
57      1.633  -.425 -1.658 -3.732 -5.176 -4.638  1.609  3.573  3.476  2.280  1.532   .612
58      1.463 -1.468 -1.012  -.957  -.440   .381  1.144  -.115   .481  1.396   .958   .189


        ANALYSIS OF STANDARDIZED RESIDUALS
               (ABSOLUTE VALUES)
 INTERVAL   NUMBER  PERCENT     CUMULATIVE
                                 PERCENT

  0 TO 1     277     39.80        39.80

  1 TO 2     160     22.99        62.79

  2 TO 3     117     16.81        79.60

 BEYOND 3    142     20.40       100.00


  SUMMARY OF FIT STATISTICS FOR ITEMS

  ITEM  AVERAGE   AVERAGE ABSOLUTE  ROOT MEAN           WEIGHTED AVERAGE  WEIGHTED ABSOLUTE
        RESIDUAL     RESIDUAL        SQUARE RESIDUAL        RESIDUAL        AVERAGE  RESIDUAL
   1     -.071        1.464            1.810                  .831             1.748
   2      .119         .723            1.083                  .476              .740
   3      .244         .781             .941                  .642              .888
   4      .117        3.112            3.726                  .114             4.181
   5      .051        1.739            2.117                  .431             2.273
   6     1.052        1.797            2.223                  .012             1.606
   7      .953        1.697            1.753                 -.074             1.448
   8      .677        3.802            4.301                -1.320             3.990
   9      .675        2.126            2.722                  .458             2.061
  10      .2[illegible]        1.596            1.961                  .870             1.987
  11      .054        1.584            1.967                 1.200             1.590
  12      .002        1.528            1.914                 1.115             1.491
  13     -.029        1.651            2.013                 1.207             1.807
  14      .099        1.256            1.505                  .882             1.223
  15     -.035         .989            1.307                  .629              .960
  16     -.233        1.126            1.336                  .146             1.182
  17      .142        1.859            2.195                 -.459             1.976
```

```
        18     -.226        .976           1.033                 -.170          1.027
        19     -.132       2.421           2.801                  .563          3.031
        20      .055       3.300           4.059                  .591          4.208
        21     -.032       3.084           1.716                  .361          3.665
        22      .060        .668            .966                 -.404           .726
        23     1.309       2.845           3.256                  .253          2.841
        24     1.027       1.884           2.322                  .073          1.509
        25     -.078       1.154           1.479                  .118          1.387
        26      .604       2.318           2.915                 -.918          2.387
        27      .314        .893           1.121                 -.100           .560
        28     1.300       4.601           4.946                 1.770          4.520
        29      .029        .920           1.085                 -.315          1.200
        30      .414       1.920           2.092                 -.141          1.857
        31      .224        .864           1.017                  .235           .945
        32      .483       2.056           2.234                 -.020          2.040
        33      .256       1.134           1.553                  .106          1.109
        34      .141        .865           1.032                 -.241          1.000
        35     1.038       2.363           2.695                 -.685          2.269
        36     1.935       7.056           7.576                 2.699          6.742
        37      .437       1.355           1.637                  .165          1.178
        38     -.053       1.631           2.954                 1.230          3.034
        39     -.737       3.374           3.505                 -.750          3.121
        40      .071       1.722           1.922                  .406          1.981
        41      .259       1.163           1.491                 -.394          1.190
        42      .233        .602            .829                  .059           .587
        43      .214        .868            .973                  .165           .834
        44     -.859       1.577           1.697                  .214          1.419
        45     -.426       1.161           1.279                  .158          1.311
        46      .815       2.006           2.128                 -.377          2.005
        47     1.853       4.31[illegible]     4.772                 2.232          4.522
        48      .914       1.493           1.947                 -.049           .915
        49     -.045       1.201           1.391                  .072          1.584
        50     -.013        .775            .363                  .264           .710
        51     1.535       3.249           3.497                 1.204          3.056
        52     -.553       5.026           5.406                 -.566          3.747
        53     -.029       1.372           1.515                 -.707          1.542
        54     -.395       1.190           1.390                  .322          1.160
        55     -.139       1.832           2.145                 -.706          2.325
        56      .066        .494            .719                  .087           .339
        57     -.025       2.479           2.858                 -.657          3.261
        58      .158        .834            .944                  .200           .671

  AVERAGES      .277       1.910           2.218                  .233          1.977


                              SUMMARY OF FIT STATISTICS FOR ABILITY LEVELS
                                       ABILITY LEVEL(MID-POINTS)
                                1       2       3       4       5       6       7       8       9      10      11      12
FIT STATISTIC                -2.75   -2.25   -1.75   -1.25    -.75    -.25     .25     .75    1.25    1.75    2.25    2.75
AVERAGE RESIDUAL               .543    .842    .660    .327    .029    .254    .398    .170   -.096    .113    .155   -.074

AVERAGE ABSOLUTE RESIDUAL     1.630   2.562   2.736   2.642   2.204   1.634   1.681   2.076   2.074   1.646   1.346    .693

ROOT MEAN SQUARE RESIDUAL     2.076   2.193   3.284   3.716   3.195   2.042   1.979   2.699   2.771   2.231   1.863   1.355
```

```
41        -.423     .757   -.[illegible]   -.132 -2.531 -1.682  -.011   -.594   2.028   3.103  1.780    .694
42        1.261     .512   -.224   -.316  -.648  -.544  -.021    .368   2.392   -.459   .231    .224
43        1.957    1.569  -1.594   -.77[illegible] -1.009  -.295   .641   1.584    .877    .534  -.332    .174
          1.[illegible]    1.222  -[illegible]   -.365   .193  2.626  1.251   -.265  -2.106  -2.519 -3.208    .232
                                                                                                    .478

OVERALL AVERAGES   AVERAGE RESIDUAL=    .277   AVERAGE ABSOLUTE RESIDUAL=   1.910   ROOT MEAN SQUARE RESIDUAL=   2.534
```

```
                              EXPECTED P VALUES-3 PARAMETER MODEL

                                                         ABILITY LEVEL
CATEGORY        1      2      3      4      5      6      7      8      9     10     11     12
MID-POINT   -2.75  -2.25  -1.75  -1.25   -.75   -.25    .25    .75   1.25   1.75   2.25   2.75
NO. OF
EXAMINEES      24     50    114    194    318    440    509    368    248     96     32     11

  ITEM
    1        .192   .294   .435   .621   .786   .894   .951   .978   .990   .990   .990   .990
    2        .452   .597   .733   .838   .908   .950   .973   .986   .990   .990   .990   .990
    3        .587   .714   .832   .904   .947   .972   .985   .990   .990   .990   .990   .990
    4        .043   .044   .048   .065   .133   .337   .671   .896   .974   .990   .990   .990
    5        .113   .125   .158   .233   .380   .590   .785   .905   .962   .985   .990   .990
    6        .175   .176   .178   .184   .200   .238   .323   .473   .666   .828   .924   .969
    7        .210   .211   .213   .219   .235   .275   .367   .530   .725   .872   .948   .980
    8        .467   .511   .555   .600   .643   .685   .725   .761   .794   .824   .851   .874
    9        .113   .119   .129   .147   .177   .225   .297   .397   .518   .644   .757   .845
   10        .151   .209   .324   .502   .697   .846   .930   .970   .988   .990   .990   .990

   11        .335   .563   .736   .916   .971   .990   .990   .990   .990   .990   .990   .990
   12        .289   .480   .701   .863   .945   .980   .990   .990   .990   .990   .990   .990
   13        .246   .415   .642   .829   .931   .974   .990   .990   .990   .990   .990   .990
   14        .293   .417   .654   .816   .914   .963   .984   .990   .990   .990   .990   .990
   15        .271   .412   .580   .740   .857   .928   .965   .983   .990   .990   .990   .990
   16        .346   .472   .613   .743   .842   .909   .949   .972   .985   .990   .990   .990
   17        .298   .316   .447   .536   .627   .711   .784   .844   .889   .923   .947   .964
   18        .012   .014   .018   .026   .041   .068   .116   .196   .316   .469   .630   .766
   19        .110   .119   .142   .204   .339   .554   .768   .902   .963   .987   .990   .990
   20        .036   .037   .042   .067   .174   .480   .820   .958   .990   .990   .990   .990

   21        .019   .020   .026   .053   .152   .415   .749   .927   .982   .990   .990   .990
   22        .675   .757   .825   .877   .916   .943   .962   .975   .983   .989   .990   .990
   23        .252   .252   .252   .253   .256   .267   .303   .407   .615   .829   .943   .983
   24        .157   .157   .159   .165   .179   .213   .289   .432   .627   .803   .911   .964
   25        .114   .124   .147   .193   .280   .418   .592   .754   .870   .936   .970   .986
   26        .363   .428   .500   .574   .647   .715   .776   .827   .869   .902   .927   .947
   27        .022   .023   .024   .027   .032   .045   .072   .129   .234   .398   .594   .766
   28        .158   .161   .166   .173   .182   .195   .214   .238   .271   .314   .366   .429
   29        .017   .019   .024   .033   .052   .088   .157   .271   .431   .610   .765   .871
   30        .148   .150   .156   .172   .213   .306   .476   .686   .850   .938   .976   .990

   31        .162   .215   .304   .435   .593   .741   .851   .921   .959   .980   .990   .990
   32        .211   .212   .217   .234   .286   .421   .650   .852   .950   .985   .990   .990
   33        .122   .135   .158   .195   .253   .338   .449   .576   .699   .802   .876   .926
   34        .779   .838   .885   .919   .944   .962   .974   .983   .988   .990   .990   .990
   35        .260   .315   .384   .463   .549   .634   .714   .784   .841   .885   .919   .943
   36        .212   .212   .212   .212   .212   .212   .212   .212   .212   .213   .217   .238
   37        .116   .125   .146   .166   .210   .278   .374   .496   .626   .745   .838   .902
   38        .115   .140   .209   .371   .624   .832   .940   .980   .990   .990   .990   .990
   39        .363   .421   .484   .551   .617   .680   .739   .790   .835   .871   .901   .924
   40        .119   .138   .180   .268   .420   .614   .787   .898   .955   .981   .990   .990
```

41	.049	.049	.052	.061	.091	.130	.234	.415	.635	.814	.918	.966
42	.151	.155	.266	.374	.514	.662	.788	.878	.933	.965	.982	.990
43	.223	.288	.395	.527	.664	.782	.868	.924	.958	.977	.987	.990
44	.174	.222	.296	.399	.525	.655	.769	.855	.914	.950	.972	.984
45	.114	.124	.143	.179	.245	.352	.498	.657	.792	.885	.940	.970
46	.073	.080	.083	.091	.114	.175	.315	.545	.772	.908	.967	.989
47	.105	.105	.105	.105	.105	.105	.106	.107	.113	.133	.202	.382
48	.042	.043	.045	.050	.059	.079	.118	.192	.315	.482	.660	.804
49	.111	.113	.137	.175	.251	.380	.556	.730	.858	.932	.969	.986
50	.264	.367	.439	.642	.768	.861	.921	.957	.977	.988	.990	.990
51	.130	.130	.131	.132	.135	.141	.156	.190	.261	.389	.569	.748
52	.406	.440	.476	.512	.550	.587	.624	.659	.694	.726	.757	.785
53	.433	.516	.603	.685	.759	.820	.870	.907	.935	.954	.969	.978
54	.172	.227	.314	.430	.584	.723	.832	.905	.949	.973	.986	.990
55	.011	.012	.015	.024	.046	.099	.214	.411	.644	.825	.925	.970
56	.196	.243	.329	.445	.579	.718	.813	.887	.935	.963	.980	.989
57	.010	.010	.010	.019	.050	.130	.298	.549	.778	.910	.967	.988
58	.177	.235	.328	.457	.605	.742	.846	.914	.954	.976	.987	.990

OBSERVED P VALUES-3 PARAMETER MODEL

	ABILITY LEVEL											
CATEGORY	1	2	3	4	5	6	7	8	9	10	11	12
MID-POINT	-2.75	-2.25	-1.75	-1.25	-.75	-.25	.25	.75	1.25	1.75	2.25	2.75
NO. OF EXAMINEES	24	50	114	194	318	440	509	368	248	90	32	11
ITEM												
1	.250	.240	.447	.613	.786	.907	.947	.981	.980	1.000	.969	1.000
2	.542	.690	.719	.835	.906	.948	.976	.986	.988	1.000	1.000	1.000
3	.792	[illegible]	.816	.897	.962	.961	.950	.995	.992	1.000	1.000	1.000
4	0.000	.080	.044	.077	.148	.345	.652	.899	.968	1.000	1.000	.909
5	.125	.140	.167	.247	.374	.584	.788	.913	.952	.978	1.000	.909
6	.083	.140	.246	.232	.182	.223	.348	.467	.645	.840	.875	1.000
7	.167	.180	.246	.237	.252	.291	.381	.503	.702	.867	.969	1.000
8	.750	.640	.605	.603	.626	.650	.676	.807	.786	.911	.938	1.000
9	.208	[illegible]	.149	.113	.182	.259	.295	.353	.520	.656	.719	.909
10	.125	.280	.368	.505	.698	.827	.923	.978	.996	.989	1.000	1.000
11	.167	.600	.807	.943	.975	.984	.996	.997	.996	1.000	1.000	1.000
12	.042	.340	.807	.892	.962	.975	.982	.995	.996	1.000	1.000	1.000
13	.083	.300	.632	.892	.925	.984	.988	.989	.992	1.000	1.000	1.000
14	.125	.460	.737	.794	.928	.966	.974	.992	1.000	1.000	1.000	1.000
15	.083	[illegible]	.614	.773	.855	.941	.967	.965	.988	1.000	1.000	1.000
16	.167	.340	.623	.784	.884	.934	.933	.948	.972	1.000	1.000	1.000
17	.500	.300	.474	.505	.613	.759	.745	.851	.895	.911	.969	1.000
18	0.000	0.000	0.000	.010	.031	.066	.132	.196	.327	.489	.406	.727
19	0.000	.080	.053	.155	.311	.623	.782	.861	.940	.989	1.000	1.000
20	0.000	.040	.044	.057	.186	.477	.819	.959	.980	1.000	1.000	1.000
21	0.000	.020	.035	.041	.145	.441	.741	.921	.972	1.000	1.000	1.000
22	.708	.760	.868	.856	.906	.939	.963	.978	.988	.989	1.000	1.000
23	.250	.220	.237	.289	.242	.310	.297	.410	.593	.789	1.000	1.000
24	.208	.120	.193	.196	.145	.225	.310	.424	.585	.789	.969	1.000
25	.083	.040	.070	.149	.296	.457	.599	.717	.863	.922	.969	1.000
26	.583	.620	.535	.531	.664	.652	.758	.864	.871	.956	1.000	1.000
27	0.000	0.000	.026	.015	.044	.056	.081	.101	.254	.333	.625	.818
28	.208	[illegible]	.184	.186	.192	.193	.204	.250	.230	.411	.313	.273
29	0.000	.040	.009	.031	.063	.105	.130	.291	.427	.567	.750	.818
30	.250	.160	.219	.160	.189	.302	.485	.674	.867	.900	1.000	1.000
31	.250	.340	.307	.412	.597	.732	.859	.913	.956	1.000	.969	1.000
32	.333	[illegible]	.246	.222	.321	.414	.654	.837	.952	1.000	.969	.909
33	.083	.180	.193	.191	.261	.330	.448	.590	.657	.789	.938	1.000
34	.958	.840	.877	.902	.947	.966	.972	.978	.992	1.000	1.000	1.000
35	.667	.560	.412	.433	.544	.582	.666	.810	.887	.956	1.000	1.000
36	.250	.160	.237	.253	.274	.230	.196	.193	.153	.178	.313	.091
37	.125	.200	.132	.196	.192	.284	.365	.514	.601	.733	.813	1.000
38	0.000	.180	.158	.387	.648	.830	.925	.981	.992	1.000	1.000	1.000
39	.292	.380	.430	.541	.645	.711	.745	.761	.831	.833	.875	.909
40	0.000	.120	.114	.294	.453	.610	.782	.908	.935	.989	1.000	1.000

41	0.000	.120	.053	.052	.060	.134	.269	.376	.621	.822	.969	1.000
42	.417	.240	.246	.392	.513	.652	.778	.875	.964	.933	1.000	1.000
43	.375	.400	.368	.536	.651	.770	.857	.946	.964	.978	.969	1.000
44	0.000	.080	.167	.397	.538	.718	.794	.845	.871	.867	.906	.909
45	0.000	.040	.061	.108	.214	.423	.509	.633	.782	.856	.906	.818
46	.157	.080	.123	.067	.388	.168	.346	.543	.746	.856	.969	1.000
47	.042	.100	.123	.124	.123	.095	.106	.114	.073	.222	.063	.455
48	0.000	.100	.038	.031	.035	.084	.110	.190	.331	.533	.469	.818
49	.042	.120	.123	.149	.239	.382	.568	.764	.819	.911	.906	1.000
50	.333	.420	.421	.675	.783	.864	.910	.951	.992	.978	1.000	1.000
51	.167	.140	.123	.129	.142	.145	.153	.196	.226	.411	.688	.545
52	.542	.560	.395	.515	.519	.605	.611	.674	.730	.733	.625	.727
53	.542	.540	.589	.686	.761	.820	.862	.905	.944	.967	.969	1.000
54	.042	.130	.254	.454	.594	.768	.835	.899	.907	.944	.969	1.000
55	0.000	0.000	.026	.065	.038	.135	.224	.427	.609	.789	.875	1.000
56	.125	.220	.351	.521	.550	.686	.817	.897	.931	.944	1.000	1.000
57	0.000	0.000	.026	.010	.047	.118	.326	.560	.742	.844	.938	1.000
58	.125	.200	.333	.433	.632	.768	.825	.899	.956	.989	1.000	1.000

```
                              RESIDUALS- 3 PARAMETER MODEL
                                 (OBSERVED-EXPECTED)

                                                  ABILITY LEVEL
               1       2       3       4       5       6       7       8       9      10      11      12

MID-POINT   -2.750  -2.250  -1.750  -1.250   -.750   -.250    .250    .750   1.250   1.750   2.250   2.750
NO. OF
EX MINEES      24      50     114     194     318     440     509     368     248      90      32      11

  ITEM
    1        .058   -.044    .012   -.008    .000    .013   -.004    .003   -.010    .010   -.021    .010
    2        .030    .083   -.014   -.063   -.002   -.002    .003    .006   -.012    .010    .010    .010
    3        .204   -.104   -.016   -.067    .015   -.010    .005    .005    .012    .010    .010    .010
    4       -.143    .036   -.034    .012    .015    .008   -.019    .003   -.006    .016    .010   -.081
    5        .012    .015    .029    .015   -.006   -.006    .013    .008   -.010   -.008    .010   -.081
    6       -.091   -.036    .068    .048   -.018   -.016    .025   -.006   -.021   -.028   -.049    .031
    7       -.044   -.131    .032    .018    .017    .016    .014   -.027   -.023   -.005    .021    .020
    8        .283    .129    .050    .063   -.018   -.035   -.049    .046   -.008    .087    .087    .126
    9        .095    .001    .020   -.034    .006    .034   -.002   -.044    .002    .011   -.038    .065
   10       -.026    .071    .044    .063    .001   -.018   -.007    .008    .008   -.001    .010    .010

   11       -.168    .037    .021    .027    .004   -.006    .096    .007    .006    .010    .010    .010
   12       -.248   -.140    .106    .028    .017   -.005   -.008    .005    .006    .010    .010    .010
   13       -.162   -.035   -.011    .062   -.007    .010   -.002   -.001    .002    .010    .010    .010
   14       -.168    .003    .083   -.022    .014    .003   -.010    .002    .010    .010    .010    .010
   15       -.196   -.152    .034    .033   -.002    .013    .002   -.019   -.002    .010    .010    .010
   16       -.179   -.132    .016    .041    .041    .025   -.016   -.024   -.013    .010    .010    .010
   17        .202   -.066    .026   -.031   -.013    .048   -.040    .007    .006   -.012    .021    .036
   18       -.012   -.014   -.018   -.016   -.009   -.002    .016   -.006    .011    .020   -.223   -.039
   19       -.113   -.039   -.090   -.049   -.028    .069    .014   -.041   -.023    .002    .010    .010
   20       -.136    .003    .002   -.010    .012   -.002   -.001    .001   -.010    .010    .010    .010

   21       -.018    .000    .009   -.012   -.007    .026   -.008   -.006   -.010    .010    .010    .010
   22        .034    .023    .043   -.022   -.010   -.005    .000    .003    .034   -.000    .010    .010
   23       -.102   -.152   -.015    .035  [.014    .033   -.006    .003   -.022   -.041    .057    .017
   24        .052   -.037    .034    .031    .034    .012    .021   -.008   -.043   -.014    .057    .036
   25       -.031   -.184   -.077   -.044    .016    .039    .008   -.037   -.007   -.014   -.011    .014
   26        .120    .192    .036   -.043    .017   -.063   -.017    .037    .002    .054    .073    .053
   27       -.022   -.123    .012   -.011    .012    .085    .038   -.028    .020   -.065    .031    .052
   28        .050   -.141    .018    .013    .010   -.002   -.009    .012   -.041    .098   -.054   -.156
   29       -.017    .021   -.015   -.062    .011    .016   -.027    .020   -.004   -.044   -.015   -.053
   30        .102    .030    .063   -.013   -.024   -.004    .010   -.012    .017   -.038    .024    .010

   31        .088    .125    .003   -.023    .005   -.009    .067   -.008   -.004    .020   -.021    .010
   32        .123    .088    .029   -.012    .035   -.008    .004   -.015    .001    .015   -.021   -.081
   33       -.039    .145    .035   -.004    .008   -.008   -.001    .014   -.042   -.013    .061    .074
   34        .181    .002   -.008   -.017    .002    .004   -.002   -.004    .014    .010    .010    .016
   35        .497    .144    .028   -.030   -.005   -.052   -.048    .026    .046    .070    .081    .057
   36        .038   -.052    .025    .041    .062    .018   -.016   -.019   -.059   -.035    .096   -.147
   37        .019    .075   -.009    .029   -.018    .007   -.009    .018   -.026   -.012   -.025    .098
   38       -.116    .140   -.051    .016    .028   -.003   -.014    .001    .002    .010    .010    .010
   39       -.071   -.041   -.155   -.009    .028    .031    .006   -.030   -.004   -.038   -.026   -.015
   40       -.119   -.018   -.066    .025    .033   -.014   -.065    .009   -.020    .008    .010    .010

   41       -.048    .071    .006   -.009   -.021    .004    .035   -.045   -.014    .008    .051    .034
```

```
42    .263   .045  -.026   .018  -.001  -.016  -.010  -.003   .031  -.031   .010   .010

43    .162   .112  -.026   .009  -.013  -.012  -.012   .021   .036   .001  -.019   .010
44   -.174  -.142  -.131  -.003   .013   .063   .025  -.016  -.043  -.083  -.065  -.075
45   -.114  -.084  -.081  -.071  -.031   .071   .011  -.023  -.010  -.029  -.034  -.152
46    .087  -.000   .040  -.024  -.025  -.006   .031  -.002  -.026  -.053   .002   .011
47   -.083  -.025   .018   .019   .018  -.010   .001   .007  -.040   .089  -.139   .072
48   -.042   .057   .043  -.019  -.025   .005  -.008  -.002   .016   .051  -.191   .014
49   -.070   .001  -.014  -.026  -.012   .001   .012   .033  -.039  -.020  -.062   .014
50    .069   .053  -.078   .033   .015   .002  -.012  -.046   .015  -.010   .010   .010

51    .037   .010  -.028  -.003   .007   .005  -.003   .066  -.035   .022   .118  -.202
52    .136   .120  -.081   .003  -.031   .018  -.013   .015   .036   .007  -.132  -.058
53    .067   .024  -.015   .001   .002   .000  -.007  -.002   .009   .012   .000   .022
54   -.130  -.127  -.060   .016   .010   .045   .002  -.006  -.041  -.028  -.017   .010
55   -.011  -.012   .011  -.019  -.008   .005   .010   .016  -.035  -.036  -.050   .030
56   -.061  -.023   .021   .075  -.028  -.021   .005   .010  -.033  -.019   .020   .011
57   -.010  -.112   .016  -.009  -.003  -.011   .028   .011  -.036  -.065  -.029   .012
58   -.052  -.035   .005  -.024   .027   .026  -.021  -.015   .032   .013   .013   .010


SUMMARY OF FIT STATISTICS FOR ITEMS

ITEM  AVERAGE   AVERAGE ABSOLUTE  ROOT MEAN             WEIGHTED AVERAGE  WEIGHTED ABSOLUTE
      RESIDUAL     RESIDUAL       SQUARE RESIDUAL           RESIDUAL        AVERAGE   RESIDUAL
 1      .002         .016              .023                   .001               .008
 2      .015         .019              .032                   .002               .006
 3      .010         .033              .066                   .001               .012
 4     -.005         .021              .030                   .001               .012
 5     -.003         .015              .025                   .000               .008
 6     -.008         .036              .042                   .001               .024
 7      .011         .022              .024                   .004               .019
 8      .058         .077              .089                   .000               .040
 9      .010         .029              .039                  -.001               .020
10      .009         .017              .025                   .001               .011
11     -.003         .026              .051                   .005               .011
12     -.017         .049              .087                   .004               .019
13     -.009         .027              .051                   .003               .012
14     -.005         .029              .055                   .003               .014
15     -.013         .032              .059                   .001               .014
16     -.018         .043              .065                   .001               .027
17      .015         .042              .065                  -.000               .029
18     -.024         .032              .062                  -.002               .013
19     -.023         .040              .046                  -.007               .038
20     -.001         .009              .013                  -.000               .005
21      .000         .011              .012                  -.000               .012
22      .009         .014              .018                   .000               .008
23     -.001         .025              .031                   .002               .019
24      .009         .032              .034                   .001               .024
25     -.018         .031              .036                  -.005               .027
26      .057         .067              .081                   .001               .038
27     -.002         .023              .029                  -.001               .015
28     -.009         .042              .060                   .000               .018
29     -.009         .020              .023                  -.001               .017
30      .014         .029              .037                   .001               .017
31      .016         .027              .043                   .001               .012
32      .013         .036              .050                   .005               .016
33      .011         .029              .035                  -.000               .014
34      .016         .021              .050                   .001               .007
35      .059         .091              .127                   .000               .046
36     -.004         .050              .062                   .002               .033
```

```
  38   -.035        .025          .041               -.002          .014
  39   -.019        .029          .030               -.001          .022
  40   -.012        .028          .041               -.002          .019
  41    .005        .028          .035               -.002          .023
  42    .025        .038          .075                .002          .015
  43    .02[illegible]        .034          .055                .001          .017
  44   -.052        .069          .070               -.003          .040
  45   -.046        .059          .057               -.008          .039
  46    .0[illegible]        .026          .035               -.002          .021
  47   -.033        .04[illegible]          .057                .001          .018
  48   -.038        .039          .062               -.002          .018
  49   -.015        .025          .029               -.003          .019
  50    .038        .026          .035                .001          .016
  51   -.034        .038          .076               -.000          .011
  52    .012        .054          .072                .001          .026
  53    .019        .013          .020                .000          .006
  54   -.027        .041          .052               -.002          .025
  55   -.038        .02[illegible]          .023               -.003          .015
  56   -.0[illegible]1        .025          .032                .000          .020
  57   -.039        .02[illegible]          .024               -.002          .019
  58   -.034        .020          .024               -.001          .019

AVERAGES  .000      .033          .046                .000          .019


                                   SUMMARY CF FIT STATISTICS FCR APILITY LEVELS
                                           ABILITY LEVEL(MID-POINTS)
                                  1       2       3       4       5       6       7       8       9       10      11      12
FIT STATISTIC                  -2.75   -2.25   -1.75   -1.25    -.75    -.25     .25     .75    1.25    1.75    2.25    2.75
AVERACE RESIDUAL                .006    .005    .001    .000    .001    .006   -.002   -.001   -.000   -.000   -.004   -.000

AVERAGE ABSOLUTE RESIDUAL       .099    .056    .034    .022    .016    .017    .012    .014    .017    .025    .039    .039

ROOT MEAN SQUARE RESIDUAL       .129    .075    .044    .028    .020    .025    .017    .019    .021    .035    .060    .059


   OVERALL AVERAGES   AVERAGE RESIDUAL=    .000   AVERAGE ABSCLUTE RESIDUAL=    .033   ROCT MEAN SQUARE RESICUAL=    .044
```

```
                              STANDARDIZED RESIDUALS- 3          PARAMETER MODEL

                                                ABILITY LEVEL
                 1       2       3       4       5       6       7       8       9      10      11      12

MID-POINT    -2.750  -2.250  -1.750  -1.250   -.750   -.250    .250    .750   1.250   1.750   2.250   2.750
NO. OF
EXAMINEES        24      50     114     134     318     440     509     368     248      90      32      11

  ITEM
   1           .723   -.696    .262   -.226    .015    .899   -.428    .349  -1.638    .953  -1.208    .333
   2           .882   1.197   -.332   -.123   -.152   -.205    .442    .079   -.332    .953    .569    .333
   3          2.034  -1.650   -.454   -.325   1.199  -1.316    .948    .880    .306    .953    .569    .333
   4         -1.042   1.230   -.222    .668    .797    .355   -.912    .207   -.582    .953    .569  -2.697
   5           .189    .310    .256    .485   -.218   -.259    .171    .527   -.840   -.593    .569  -2.697
   6         -1.179   -.661   1.887   1.713   -.791   -.774   1.216   -.225   -.690   -.714  -1.048    .594
   7          -.526   -.541    .840    .601    .698    .741    .679  -1.047   -.827   -.141    .529    .471
   8          2.790   1.830   1.076    .096   -.652  -1.588  -2.464   2.064   -.320   2.162   1.375   1.259
   9          1.479    .025    .633  -1.323    .260   1.726   -.123  -1.713    .066    .220   -.507    .591
  10          -.351   1.228   1.012    .097    .340  -1.070   -.587    .917   1.192   -.106    .569    .333

  11         -1.747    .521    .538   1.360    .450  -1.246   1.377   1.404    .945    .953    .569    .333
  12         -2.676  -1.476   2.463   1.155   1.324   -.671  -1.742    .880    .945    .953    .569    .333
  13         -1.849   -.496   -.225   2.312   -.474   1.280   -.405   -.168    .316    .953    .569    .333
  14         -1.812    .036   1.868   -.790    .871    .367  -1.776    .356   1.583    .953    .569    .333
  15         -2.141   -.751    .743   1.061   -.096   1.469    .195  -2.825   -.332    .953    .569    .333
  16         -1.845  -1.870    .215   1.298   2.015   1.826  -1.671  -2.829  -1.755    .953    .569    .333
  17          2.164   -.976    .564   -.371   -.492   2.227  -2.172    .363    .288   -.430    .543    .639
  18          -.546   -.853  -1.461  -1.379   -.832   -.145   1.122   -.013    .362    .375  -2.616   -.305
  19         -1.721   -.843  -2.745  -1.705  -1.038   2.904    .729  -2.628  -1.956    .187    .569    .333
  20          -.950    .104    .092   -.549    .547   -.097   -.036    .073  -1.608    .953    .569    .333

  21          -.672    .032    .582   -.725   -.368   1.102   -.438   -.451  -1.218    .953    .569    .333
  22           .351    .278   1.221   -.924   -.665   -.426    .058    .406    .545   -.824    .569    .333
  23          -.023   -.848   -.381   1.137   -.573   1.576   -.292    .133   -.704  -1.023   1.387    .432
  24           .639   -.725    .978   1.161  -1.590    .625   1.054   -.327  -1.389   -.337   1.141    .645
  25          -.471  -1.809  -2.317  -1.541    .538   1.659    .349  -1.644   -.310   -.536   -.041    .392
  26          2.246   2.743    .760  -1.208    .619  -2.922   -.934   1.887    .103   1.712   1.582    .786
  27          -.742  -1.083    .158   -.965   1.176    .501    .700  -1.627    .739  -1.255    .356    .409
  28           .678   -.793    .522    .471    .441   -.122   -.516    .523  -1.464   1.996   -.631  -1.048
  29          -.644   1.076  -1.044   -.150    .916   1.199  -1.673    .867   -.122   -.850   -.199   -.527
  30          1.402    .584   1.847   -.463  -1.056   -.171    .433   -.492    .741  -1.513    .880    .333

  31          1.168   2.159    .072   -.646    .170   -.428    .456   -.540   -.300   1.366  -1.205    .333
  32          1.475   1.122    .741   -.394   1.381   -.331    .189   -.869    .087   1.173  -1.208  -2.697
  33          -.513    .922   1.035   -.133    .343   -.360   -.052    .534  -1.436   -.301   1.050    .938
  34          2.128    .039   -.251   -.889    .163    .423   -.251   -.639    .533    .953    .569    .333
  35          4.544   3.715    .614   -.848   -.167  -2.286  -2.403   1.217   1.997   2.093   1.683    .814
  36           .455   -.900    .649   1.383   2.687    .900   -.858   -.896  -2.270   -.814   1.313  -1.145
  37           .135   1.607   -.264   1.102   -.789    .305   -.410    .687   -.830   -.258   -.391   1.092
  38         -1.779    .805  -1.351    .452   1.033   -.147  -1.373    .095    .306    .953    .569    .333
  39          -.727   -.589  -1.166   -.263   1.017   1.390    .295  -1.394   -.163  -1.067   -.484   -.188
  40         -1.798   -.361  -1.840    .800   1.191   -.608   -.294    .594  -1.496    .549    .569    .333
```

```
41      -1.098  2.322   .013  -.531 -1.401   .251  1.048 -1.756  -.457   .195  1.649   .621
42       3.509   .032  -.499   .517  -.043  -.425  -.564  -.159  1.929 -1.802   .776   .333
43       1.937  1.756  -.572   .250  -.563  -.596  -.790  1.547   .463   .057  -.951   .333
44      -2.249 -2.419 -3.031  -.173   .452  2.781  1.312  -.564 -2.393 -3.625 -2.223 -1.981
45      -1.759 -1.799 -2.430 -2.571 -1.287  3.120   .503  -.946  -.371  -.873  -.805 -2.936
46       1.580  -.039  1.539 -1.158 -1.433  -.356  1.496  -.059  -.990 -1.738   .055   .356
47      -1.812  -.115   .620   .850  1.324  -.663   .037   .434 -2.005  2.477 -1.964   .493
48      -1.024  1.993  2.198 -1.196 -1.854   .408  -.580  -.097   .538   .967 -2.283   .121
49      -1.086   .013  -.448  -.949  -.490   .058   .543  1.444 -1.764  -.767 -2.026   .395
50        .765   .782 -1.672   .960   .623   .137  -.996  -.562  1.574  -.860   .569   .333

51        .532   .203  -.253  -.126   .361   .272  -.162   .296 -1.249   .426  1.351 -1.545
52       1.352  1.005 -1.735   .083 -1.104   .753  -.587   .588  1.232   .147 -1.741  -.469
53        .662   .333  -.323   .017   .098   .002  -.473  -.137   .572   .556   .007   .493
54      -1.692 -2.137 -1.377   .437   .376  2.180   .151  -.372 -2.939 -1.647  -.811   .333
55       -.512  -.782   .943 -1.720  -.711   .369   .543   .616 -1.161  -.911 -1.079   .583
56       -.769  -.382   .487  2.110 -1.028  -.984   .262   .583  -.206  -.942   .818   .354
57       -.492  -.711  1.751  -.915  -.260  -.708  1.395   .422 -1.348 -2.151  -.914   .364
58       -.665  -.188   .114  -.679   .979  1.239 -1.334 -1.008   .133   .810   .639   .333


        ANALYSIS OF STANDARDIZED RESIDUALS
                  (ABSOLUTE VALUES)
 INTERVAL   NUMBER  PERCENT       CUMULATIVE
                                    PERCENT

   0 TO 1     457     65.66         65.66

   1 TO 2     181     26.01         91.67

   2 TO 3      52      7.47         99.14

   BEYOND 3     6       .86        100.00



  SUMMARY OF FIT STATISTICS FOR ITEMS

  ITEM  AVERAGE   AVERAGE ABSOLUTE   ROOT MEAN            WEIGHTED AVERAGE   WEIGHTED ABSOLUTE
        RESIDUAL     RESIDUAL        SQUARE RESIDUAL           RESIDUAL         AVERAGE   RESIDUAL
   1     -.053         .642              .779                 -.028               .584
   2      .275         .467              .512                  .067               .302
   3      .290         .914             1.018                  .269               .916
   4     -.056         .853             1.062                  .038               .613
   5     -.174         .594              .873                 -.013               .383
   6     -.036         .958             1.060                  .070               .908
   7      .123         .637              .661                  .202               .738
   8      .635        1.472             1.553                 -.387              1.481
   9      .111         .722              .951                  .012               .817
  10      .273         .625              .705                  .032               .688
  11      .455         .953              .950                  .610              1.102
  12      .133        1.307             1.486                  .103              1.225
  13      .178         .781             1.007                  .284               .713
  14      .213         .943             1.113                  .077               .995
  15     -.102         .922             1.201                 -.114               .921
  16     -.231        1.432             1.597                 -.265              1.790
  17      .154         .977             1.206                 -.080              1.186
  18     -.524          .834              .955                      -.090              .669

  19     -.653         1.446             1.585                      -.343             1.750
  20     -.047          .493              .673                      -.110              .381
  21     -.327          .618              .700                      -.127              .677
  22      .152          .492              .567                      -.032              .452
  23      .358          .709              .854                       .137              .707
  24      .151          .889              .953                       .072              .936
  25     -.469          .976             1.122                      -.117             1.047
  26      .614         1.458             1.568                      -.245             1.419
  27     -.136          .809              .897                       .082              .892
  28      .015          .767              .911                      -.030              .596
  29     -.036          .772              .888                       .024              .970
  30      .211          .826              .944                      -.045              .637
  31      .217          .737              .918                      -.029              .493
  32      .034         1.000             1.222                       .113              .603
  33      .153          .641              .743                      -.039              .468
  34      .259          .598              .757                      -.014              .457
  35      .914         1.865             2.039                      -.370             1.670
  36      .042         1.189             1.346                       .077             1.295
  37      .155          .656              .764                      -.015              .595
  38     -.009          .766              .925                      -.119              .682
  39     -.279          .729              .807                       .078              .827
  40     -.137          .869             1.305                      -.098              .785
  41      .088          .961             1.191                      -.045             1.068
  42      .388          .935             1.281                      -.005              .617
  43      .244          .813              .973                      -.018              .749
  44    -1.168         1.925             1.858                       .117             1.580
  45    -1.017         1.621             1.565                      -.113             1.472
  46     -.050          .897             1.107                      -.117              .935
  47      .014          .975             1.227                       .036              .772
  48     -.058         1.105             1.343                      -.208              .835
  49     -.423          .832              .945                      -.062              .762
  50      .138          .819              .915                      -.028              .791
  51      .009          .565              .746                      -.006              .380
  52      .019          .958             1.119                       .038              .796
  53      .150          .306              .355                      -.026              .247
  54     -.632         1.198             1.341                      -.057             1.065
  55     -.318          .928              .842                      -.098              .749
  56      .025          .744              .893                      -.038              .741
  57     -.297          .953             1.060                      -.045              .936
  58     -.032          .710              .803                      -.095              .928

AVERAGES -.002          .905             1.042                      -.020              .858



                                      SUMMARY OF FIT STATISTICS FOR ABILITY LEVELS
                                            ABILITY LEVEL (MID-POINTS)
                                 1        2        3        4        5        6        7        8        9       10       11       12
FIT STATISTIC                -2.75    -2.25    -1.75    -1.25     -.75     -.25      .25      .75     1.25     1.75     2.25     2.75
AVERAGE RESIDUAL              -.039     .079     .049    -.065     .066     .270    -.152    -.085    -.344     .099     .052     .045

AVERAGE ABSOLUTE RESIDUAL     1.278    1.022     .961     .844     .758     .922     .790     .809     .947     .964     .851     .674

ROOT MEAN SQUARE RESIDUAL     1.537    1.287    1.220    1.020     .923    1.184     .990    1.059    1.116    1.168    1.051     .940



   OVERALL AVERAGES    AVERAGE RESIDUAL=   -.002   AVERAGE ABSOLUTE RESIDUAL=    .905   ROOT MEAN SQUARE RESIDUAL=   1.125
```